A New Closed-Form Two-Stage Budgeting-Based Multiple Discrete-Continuous Model

Chandra R. Bhat
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, USA
Tel: 1-512-471-4535; Email: bhat@mail.utexas.edu
and
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ABSTRACT

In this paper, we propose a multiple discrete-continuous (MDC) model approach that (a) does not need the total budget to be observed or predetermined, (b) allows for any finite or not-so-finite budget over the entire set of inside and outside goods, and (c) preserves a strong endogenous utility-theoretic link between inside good consumptions and the budget allocated to the inside goods (that is, to the product group of interest). We show that our proposed model, including a fractional MDC model at the lower level linked up to a Tobit model for the budget allocation to the inside goods, is strictly consistent with a two-stage budgeting utility theoretic structure. As importantly, by using reverse Gumbel distributional assumptions for the stochastic terms in the model system, we derive an incredibly simple closed-form model that, to our knowledge, is a first of its kind in the econometric literature. In doing so, we formally introduce a new distribution, which we label as the minLogistic distribution, to the statistical literature, and derive the properties of the distribution that is then used in the forecasting stage of the proposed model. An application of the proposed model to investigate the household vehicle fleet composition and usage demonstrates its potential relative to an unlinked and exogenously developed budget for the inside goods. The proposed model has the potential to open up a whole new world of MDC applications in general, and particularly for those cases with an unobserved total budget over the inside and outside goods.

Keywords: Multiple Discrete-Continuous Model, Two-Stage Budgeting, Reverse Gumbel Distribution, Utility-Theoretic Model, MinLogistic Distribution, Vehicle Fleet Modeling

1. INTRODUCTION

There has been a surge in the application of discrete-continuous models in several fields in recent years, particularly for the case of multiple discrete-continuous (or MDC) choice situations. Such situations are characterized by the choice of consuming multiple alternatives from a set of mutually-exclusive alternatives (rather than only one of the mutually-exclusive alternatives, as in traditional "single" discrete choice models), along with a continuous non-zero consumption intensity for each of the consumed alternatives (see Bhat, 2008). These MDC choices are pervasive in the social sciences, including the transportation, recreation, economics, marketing, actuarial science, ecology, biostatistics, and epidemiology fields. Examples include individuals' time-use choices (decisions to engage in different types of activities and time allocation to each activity), investment portfolios (where and how much to invest), and grocery purchases (brand choice and purchase quantity).

Three broad approaches in the literature to model MDC data include those associated with (1) statistical censoring, (2) vertical choice-making behavior, and (3) horizontal choice-making behavior. The first approach, based on Tobin's (1958) model (see, for example, Srinivasan and Bhat, 2006, Fang, 2008, Anastasopoulos et al., 2012, and Hou et al., 2020), uses a set of latent variables to represent continuous demand for each of the alternatives, and then uses a censoring mechanism to allow for zero consumption (that is, the corresponding alternative is not chosen for consumption) and positive consumptions (the corresponding alternative is chosen for consumption, and the latent variable value, which is positive, represents the continuous consumption intensity). Interestingly, while this multivariate Tobit-based approach (based on the classical multivariate Tobit model and its extended variants) has typically been viewed as a reduced-form statistical stitching approach that is not utility-theoretic, Saxena et al. (2022a) recently showed that this approach can in fact be viewed as a very restrictive case of a utility-based model that assumes no differential satiation effects across inside goods and has an infinite budget available for consumption. The approach also can get tedious and be fraught with computational instability as the number of alternatives increases.

The second approach, the vertical choice-making behavior approach, assumes that the multiple discreteness is a consequence of a "vertical" stream of multiple choice occasions of an individual, and the individual is assumed to choose one and only one alternative at each choice instance (that is, at each choice instance, a single discrete choice model is assumed to apply). However, preferences may vary within the same individual across choice instances, leading to the observation of multiple discreteness over the vertical stream of choice instances (that is, one and only one alternative is observed to be consumed at each choice occasion; then, across multiple single choice occasions, one observes different alternatives being chosen). Also, at each choice occasion, the vector of goods under consideration may exhaust the consumption space of consumers across all product categories (that is, represent a complete demand system) or may focus only on goods within a specific product category (that is, represent an incomplete demand system). The complete demand system requires data on prices and consumptions of all commodity/service items, and can be impractical when studying consumptions in finely defined commodity/service categories (see Bhat and Pinjari, 2014). Thus, it is rather ubiquitous to use an incomplete demand system. Within this incomplete demand vertical behavior approach, which has been more often applied to multiple discrete-count models rather than multiple discrete-continuous models (that is, the budget is more often in the form of counts rather than a continuous value), the product of the probability of choice of an alternative at any choice instance with the budget provides the budget allocation to each good in the product category under consideration. The budget (for the product category under consideration) itself is either exogenously determined (no linking between the instance-specific single choice models at each instance and the budget) or endogenously determined (linking present between the single choice model and the budget). Examples of the former exogenous budget method include Morey et al., 1993, Hendel, 1999, and Paleti et al., 2014. This exogenous budget method, while easy to estimate and implement, does not explicitly consider the substitution and income effects that are likely to lead to a change in the budget because of a change in a variable that impacts the single discrete choice. The structure without this linkage is also not consistent with utility theory (see Bhat et al., 2015a). In the latter endogenous linking method, the single discrete choice model corresponding to choice instances is linked to the budget. Examples of the use of this method include Bockstael et al., 1987, Mannering and Hamed, 1990, Hausman et al., 1995, Rouwendal and Boter, 2009, Bhat et al., 2015a, and Wagner et al., 2019). When the single discrete choice for alternatives within a product category at any choice instance is based on the typically used multivariate extreme-value error terms (such as based on logit or nested logit or other GEV models), Rouwendal and Boter (2009) show that the use of the expected maximum utility from the discrete choice model as the price index for the product group as the linkage to the budget equation for that product group is consistent with a two-stage budgeting utility maximizing framework if the linkage is motivated from a separable indirect utility function approach across product groups. But all these methods and strategies of this second vertical choice making approach assume that the multiple discrete-continuous observations are a result of repeated single discrete-continuous (SDC) choice occasions. Such a vertical choice behavior process based on the aggregate accumulation of single discrete choice decisions at each of the many event instances, does not consider the possibility that the multiple discrete-continuous observations may originate from a single horizontal choice of multiple alternatives simultaneously, as discussed next.

The third approach, the “horizontal choice-making behavior” approach, considers the multiple discreteness as arising fundamentally from inherently imperfect substitutes at a single choice occasion (see Wales and Woodland, 1983, Kim et al., 2002, von Haefen and Phaneuf, 2003, and Bhat, 2005). This approach is the one of interest in the current paper, and so will be discussed at some length here. Basically, in many choice situations, each decision agent may choose multiple alternatives (from the set of available alternatives) horizontally and simultaneously. For example, in the recreational literature, the interest may be on the annual expenditure on fishing trips, and the split of this annual expenditure to different angler sites based on site amenities, fish catch rates, fish size, target species, bag costs, trip distance, and angler characteristics. Similarly, in the transportation field, the focus may be on the time spent in leisure activities over a weekend day and the split of this time across different leisure activity types based on built environment characteristics, activity participation costs, and individual/household characteristics. In these choice situations, it is only reasonable to assume that satiation (or variety-seeking) effects set in as the intensity of investment (of expenditures and times) in any single alternative increases. In other words, consumers horizontally select an assortment of goods due to diminishing marginal effects (that is, satiation or variety-seeking effects) for each good rather than as a vertical collection of multiple single choice instances. In this horizontal choice-making behavior approach, consumers are assumed to maximize a direct utility function $U(\mathbf{x})$ over a set of non-negative consumption quantities $\mathbf{x}=(x_1,...,x_k,...,x_K)$ subject to a budget constraint, as below:

$$Max\ U(\mathbf{x}) \text{ such that } \mathbf{x.p}=E \text{ and } x_k \geq 0, \tag{1}$$

where $U(\mathbf{x})$ is a quasi-concave, increasing and continuously differentiable non-linear utility function with respect to the consumption quantity vector, $\mathbf{p}$ is the vector of unit prices for all goods, and E is the total expenditure (or income). Again, it is common to use an incomplete demand system, typically in the form of the use of a Hicksian composite commodity approach in which the analyst assumes that the prices of elementary goods within each broad product group of consumption items vary proportionally. This Hicksian composite group approach is discussed more in the next section.

1.1. The Hicksian Composite Group Approach

The Hicksian approach works by replacing all the elementary alternatives within each broad group (that is not of primary interest) by a single composite alternative representing the broad group. The analysis proceeds by considering the composite goods as “outside” goods and considering consumption in these outside goods as well as the “inside” goods representing the product group of main interest to the analyst. It is common in practice in this Hicksian approach to include a single outside good (considered essential in that there is some positive consumption of this essential good) with the inside goods (see von Haefen, 2010). Generally, the outside good is treated as a numeraire with unit price, implying that the prices and characteristics of all goods grouped into the outside category do not influence the expenditure allocation among the inside goods (see Deaton and Muellbauer, 1980). The outside good allows for the overall demand for the inside goods to change due to changes in prices and other influential factors of the inside goods. In this way, the Hicksian approach may be viewed as a single stage utility-theoretic approach, where changes in the price (or other attribute) of any inside good lead to reallocations between the outside good and the many inside goods. Typically, within this Hicksian approach, a direct utility approach is taken to solving the constrained utility maximization problem of Equation (1), in which the utility function $U(\mathbf{x})$ is considered to be random over the population. Then, applying the Karush-Kuhn-Tucker (KKT) first-order conditions, one can derive the probabilities for any consumption pattern (including corner solutions; see Wales and Woodland, 1983, Kim et al., 2002, von Haefen and Phaneuf, 2003, and Bhat, 2005).

Within the Hicksian direct utility approach, a number of different utility forms for $U(\mathbf{x})$ have been used in the literature for the MDC case. Most of these assume additive separability of preferences in the utility form (but see Vasquez-Lavin and Hanemann, 2008 and Bhat et al., 2015b for relaxations of this assumption). Bhat (2008) proposed a Box-Cox utility function form for the inside good utilities and a non-linear utility form for the outside good utility that is quite general and subsumes earlier utility specifications as special cases, and that is consistent with the notion of weak complementarity (Mäler, 1974), which implies that the consumer receives no utility from a non-essential good's attributes if she/he does not consume it. Then, if a multiplicative log-extreme value error term is superimposed to accommodate unobserved heterogeneity in the baseline utility preference for each alternative, the result is the MDCEV model, which has a closed-form probability expression and collapses to the MNL in the case that each (and every) decision-maker chooses only one alternative. Alternative error term specifications in the baseline utility preference have also been used, leading to the MDC probit (MDCP) model and finite mixture MDCP models (for example, Bhat et al., 2013, Bhat et al. 2016, and Saxena et al., 2022b).

Regardless of the error distribution assumption made, the application of the Hicksian composite approach with non-linear utility forms for the inside and outside goods necessarily requires the budget *E* to be observed. In recent years, a variant of Bhat's (2008) utility form has received increasing attention (Bhat, 2018, Bhat et al., 2020, and Saxena et al., 2022a), based on employing a linear utility structure for the outside good. Such a linear outside good utility structure has the advantage of not needing observation of the budget quantity, which indeed may be unobserved in many situations.[1] This utility variant also will generally provide better discrete component predictions than the traditional non-linear outside good utility form, especially when the outside good takes up a substantial share of the continuous consumption.[2] However, as indicated in Saxena et al. (2022a), this linear outside good utility form may not work well in terms of data fit and prediction ability when the overall budget amount, even if unobserved, is known to be non-infinite and the outside good takes up only a small share of the continuous consumption. The reason is that the formulation, while ensuring the positivity of consumptions of the inside goods (that may or may not be consumed), does not guarantee, within the model formulation and estimation itself, the positivity of the consumption of the essential outside good.

[1] Some earlier studies have tried to "skirt" the overall budget unobservability problem by imposing a natural maximum constant budget amount across all individuals, or adding a constant allocation to an outside good to construct an overall budget, or developing a customized maximum budget amount for each individual, or developing an independent first-stage regression model for total budget (see, for example, Bhat and Sen, 2006, Pinjari et al., 2016). But these are all rather *ad hoc* ways of developing a total budget.

[2] In the traditional MDC model estimation, a large budget expenditure on the outside good will tend to drive the baseline preferences of the inside goods to small values and also drive the satiation to be extremely high for these goods, resulting in convergence instability. On the other hand, the use of a linear utility form for the outside good, because it focuses better on fitting the discrete probabilities and does not involve the appearance of the outside good consumption in the baseline preference for the inside goods handles such situations much better.

In a recent paper, Bhat et al. (2022) address this situation by making a truncation correction to ensure positivity of the outside good consumption in estimation when the budget is known or a finite limit can be placed to the budget even if unknown. This improves the accuracy of model parameter estimation and the resulting predictions. There are, however, two problems with this study. First, the overall budget needs to be known or needs to be set. In some cases, the ceiling on the budget may be known, such as the number of hours in a day in a time-use model. But this is not always the case, and ad hoc assumptions will need to be made. Second, the study, because it is based on a Hicksian composite good approach, does not expressly consider potential exogenous variable effects on an overall budget that can then impact individual good consumptions. That is, by defining the goods of interest as inside goods, changes in exogenous variables directly impact the consumptions of these inside goods (even if the true effect is an indirect impact through budget changes), co-mingling strict budget effects (that is, income effects) and strict allocation effects (that is, substitution effects). For example, consider the case of a price decrease of a particular inside good. This will have a substitution effect in terms of a draw away from other inside goods to the good whose price was lowered. However, the overall consumption on the product group (that comprises all the inside goods) may also increase because the overall group price index has now dropped, leading to an increase in the consumption of each inside good because of an income effect. The net effect on consumption of the inside goods will be a combination of the income and substitution effects, which is not handled by the Hicksian composite good approach.[3] As stated by Bhat et al. (2020), "approaches to handle both an endogenous budget as well as consumption quantities separately but within a single unifying utility-theoretic framework have been elusive; additional investigations in this area are certainly an important direction for further research."

1.2. The Current Paper in Context

In this paper, we develop a new framework for MDC models based on a utility-consistent two-stage budgeting approach (rather than the single stage Hicksian composite good approach), which can handle unobserved but finite and endogenous budgets. Our approach is based on an endogenous linking of a fractional split MDCEV model with a total budget equation for the specific product group (that contains the inside goods) under consideration. This approach, a first to our knowledge in the econometric literature, is also different from earlier MDC efforts that focus exclusively on the set of inside goods and that consider the budget for the inside goods as being determined exogenously (as in Bhat, 2005). In such earlier efforts, the lack of any linking between the inside good allocations and the total budget for the inside goods ignores the substitution and income effects that are likely to lead to a change in the budget because of a change in a variable that impacts any single inside good choice. In our endogenous linking, we ensure compatibility with a two-stage budgeting utility maximizing structure by invoking a separable direct utility function between the product category of interest and other product

[3] This is also true of the Hicksian composite good-based stochastic-frontier approach of Augustin et al. (2015) and Pinjari et al. (2016).

categories. In doing so, we develop an appropriate and economic theory-consistent linking function (which is really a price index for the product group under consideration that can represent the entire product group at the higher level of budget allocation to the product group, followed by the second stage fractional allocation to each good within the product group). Specifically, we start from Bhat's (2008) utility formulation for the second stage fractional allocation within a single product group (with the budget of the fractions across all alternatives in the product group being equal to one), and adopt a reverse Gumbel distributional assumption for the stochastic terms in the baseline preferences of each of the inside alternatives. We also consider a reverse Gumbel distribution for the random error term for the total budget for the product category in the first stage, including a stochastic group price index for the product category as developed from the second stage as an exogenous stochastic variable. By including censoring in the budget equation for the product category of interest, we accommodate the possibility of zero allocation to the product category. With these assumptions on the statistical distribution of the error terms, we formally introduce a new univariate distribution to the statistical literature, which we label as the "minLogistic" distribution, and derive its properties and moments. The end-result is an incredibly surprising closed-form model for the resulting multiple discrete-continuous extreme value model that, to our knowledge, is a first of its kind in the econometric literature.

2. MODEL FORMULATION AND STATISTICAL SPECIFICATION

2.1. Reverse Gumbel MDCEV (RG-MDCEV) Model of Fractional Split

In this section, we start with the second stage of the allocation among inside goods within the product category under consideration (say product category *G*). We use the MDCEV framework as a fractional allocation model, with the fractional allocation based on consumption expenditures rather than consumption quantities of the inside goods. This facilitates the use of the MDCEV model as a conditional (on total expenditure on the product category) multiple discrete expenditure fractional allocation. As we show later, the use of expenditure allocation (rather than quantity allocation) at the second stage of a two-stage utility-theoretic approach is necessary for consistency with a Gorman polar utility form-based linking with the first stage.

Consider the following constrained direct utility form:

$$U(\tilde{\mathbf{f}}) = \sum_{k=1}^{K} \gamma_k \psi_k \ln\left\{\left(\frac{\tilde{f}_k}{\gamma_k}+1\right)\right\} \tag{2}$$

$$s.t.\ \sum_{k=1}^{K} \tilde{f}_k = 1,$$

where $\tilde{f}_k = p_k x_k / y$, and y is the total budget allocated to product category *G*. An important point to note is that the utility function of Equation (2) is written in the form of continuous fractional splits (allocations) of consumption of the total non-zero budget (y>0) allocated to the product category of interest, such that the fractions sum to one conditional on a non-zero allocation to the product category. The possibility of zero allocation to the product category

(y =0) is handled through a censoring mechanism in the total budget for the product category, as discussed later. Also, in the above utility function, $U(\tilde{\mathbf{f}})$ is a quasi-concave, increasing, and continuously differentiable function with respect to the fractional consumption quantity (K×1)-vector $\tilde{\mathbf{f}}$ ($0 \le \tilde{f}_k \le 1$ for all k), and ψ_k and γ_k are parameters associated with good k.[4] The function $U(\tilde{\mathbf{f}})$ in Equation (2) is a valid utility function if $\psi_k > 0$, and $\gamma_k > 0$ for all k (we will use the terms "good" and "alternative" interchangeably to refer to any good k). As discussed in detail in Bhat (2008), ψ_k represents the baseline marginal utility, and γ_k is the vehicle to introduce corner solutions (that is, zero fractional splits) for the goods, but also serves the role of a satiation parameter (higher values of γ_k imply less satiation). The satiation operates at the fractional split level, so that the marginal utility of a good decreases as the fractional split investment in the good increases.

To ensure the non-negativity of the baseline marginal utility, while also allowing it to vary across individuals based on observed and unobserved characteristics, ψ_k is parameterized as follows:

$$\psi_k = \exp\left(\boldsymbol{\beta}'\mathbf{z}_k - \frac{1}{\sigma}\ln p_k + \varepsilon_k\right),\ k = 1, 2, ..., K, \tag{3}$$

where $\mathbf{z}_k$ is a set of attributes that characterize alternative k and the decision maker (including a constant), p_k is the unit price for good k, the inverse of σ (σ>0) is the coefficient on $\ln p_k$, and ε_k is a standardized scale error tern that captures the idiosyncratic (unobserved) characteristics that impact the baseline utility of good k. A constant cannot be identified in the $\boldsymbol{\beta}$ term for one of the K alternatives. Similarly, individual-specific variables are introduced in the vector $\mathbf{z}_k$ for (K–1) alternatives, with the remaining alternative serving as the base.[5] Also, the parameter σ is estimable if there is price variation across the inside goods (and may be normalized to one when there is no price variation). For later use, we will write the baseline marginal utility ψ_k in Equation (3) as:

$$\psi_k = \mu_k \exp\left(\varepsilon_k\right),\ \text{where}\ \mu_k = \exp(\boldsymbol{\beta}'\mathbf{z}_k)\left(\frac{1}{p_k^{1/\sigma}}\right). \tag{4}$$

[4] The assumption of a quasi-concave utility function is simply a manifestation of requiring the indifference curves to be convex to the origin (see Deaton and Muellbauer, 1980, p. 30 for a rigorous definition of quasi-concavity). The assumption of an increasing utility function implies that for any good k, the subutility function is such that $U_k(\tilde{f}_k^1) > U_k(\tilde{f}_k^0)$ if $\tilde{f}_k^1 > \tilde{f}_k^0$.

[5] The origin of these identification conditions is the unit sum constraint associated with the fractional splits. Also, note that one could as well have considered the baseline marginal utility function for alternative k in Equation (3) as: $\psi_k = \exp\left(\vec{\boldsymbol{\beta}}'\mathbf{z}_k - \ln p_k + \vec{\varepsilon}_k\right)$, with the error term $\vec{\varepsilon}_k$ being distributed with scale σ. It is easy to show that nothing changes in our entire formulation or model results, except that $\vec{\boldsymbol{\beta}} = \boldsymbol{\beta}\sigma$. We prefer to use the notation in Equation (3) because it simplifies the presentation.

μ_k corresponds to the systematic part of the baseline utility of alternative *k*. In the current paper, and unlike Bhat (2008) and many other MDCEV applications, we will assume a standard extreme value type 1 (Gumbel) distribution based on the limiting distribution of the minimum of random variables (that is, a standardized reverse Gumbel specification) for the ε_k terms (the reason for this will become clear later):

$$f_{\varepsilon_k}(t) = e^{-e^t}.e^t \text{ and } F_{\varepsilon_k}(t) = \text{Prob}(\varepsilon_k < t) = 1 - e^{-e^t} \text{ for } k = 1,2,3,...,K. \tag{5}$$

A similar reverse Gumbel distribution has been assumed in Mondal and Bhat (2021) and Bhat et al. (2022), though for quite different reasons than the motivation here. Based on the above reverse Gumbel distribution form for each error term, one can write the joint multivariate survival distribution function (SDF) for the error terms $\tilde{\eta}_k = \varepsilon_k - \varepsilon_1$ as follows:[6]

$$\boldsymbol{S}_{\tilde{\boldsymbol{\eta}}}(t_2,t_3,...,t_K) = \text{Prob}(\tilde{\eta}_2 > t_2, \tilde{\eta}_3 > t_3, ..., \tilde{\eta}_K > t_K) = \frac{1}{\left(1+\sum_{k=2}^{K} e^{t_k}\right)}. \tag{6}$$

The multivariate cumulative distribution function (CDF) of the $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_2,...,\tilde{\eta}_K)$ vector can be written as a function of the SDFs corresponding to the random variates as follows:

$$\boldsymbol{F}_{\tilde{\boldsymbol{\eta}}}(t_2,t_3,...,t_K) = \text{Prob}(\tilde{\eta}_2 < t_2, \tilde{\eta}_3 < t_3, ..., \tilde{\eta}_K < t_K) = 1 + \sum_{D \subset \{2,...K\}, |D| \geq 1} (-1)^{|D|} \boldsymbol{S}_D(\mathbf{t}_D), \tag{7}$$

where $\boldsymbol{S}_D(.)$ is the SDF of dimension *D*, *D* represents a specific combination of the $\tilde{\eta}$ terms (representing a specific sub-vector of the $\tilde{\boldsymbol{\eta}}$ vector; there are a total of $(K-2)+C(K-2,2)+C(K-2,3)+...C(K-2,K-2) = 2^{K-2}-1$ possible combinations, |*D*| is the cardinality of the specific combination *D*, and $\mathbf{t}_D$ is a sub-vector of the vector $\mathbf{t} = (t_2,t_3,...,t_K)$ with the appropriate elements corresponding to the combination *D* extracted. The probability density function corresponding to Equation (6) and Equation (7) may be derived in a straightforward manner as below:

$$\begin{aligned} f_{\tilde{\boldsymbol{\eta}}}(t_2,t_3,...,t_K) &= (-1)^{|K-1|} \frac{\partial S(t_2,t_3,...,t_K)}{\partial t_2 \partial t_3 ... \partial t_K} = (-1)^{|K-1|}(-1)^{|K-1|} \left| \frac{\partial S(t_2,t_3,...,t_K)}{\partial t_2 \partial t_3 ... \partial t_K} \right| \\ &= (K-1)! \left(\frac{\exp(\sum_{k=2}^{K} t_k)}{\left[1+\sum_{k=2}^{K} \exp(t_k)\right]^K} \right) \end{aligned} \tag{8}$$

[6] The $\tilde{\eta}_k$ error terms (*k* = 2, 3,…, *K*) are essentially multivariate logistically distributed with a correlation of 0.5, with the SDF expression as given below (see Bhat et al., 2020).

To find the optimal fractional splits of goods, the Lagrangian is constructed and the first order equations are derived based on the Karush-Kuhn-Tucker (KKT) conditions. The procedure is identical to Bhat (2008), except with a different error distribution than in Bhat (2008). Specifically, the Lagrangian function for the model, when combined with the budget constraint, is:

$$L = U(\tilde{\mathbf{f}}) + \lambda\left(1 - \sum_{k=1}^{K} \tilde{f}_k\right), \tag{9}$$

where λ is a Lagrangian multiplier for the constraint. The KKT first-order conditions for optimal fractional allocations ($\tilde{f}_k^{op}$) are as follows:

$$\psi_k \left(\frac{\tilde{f}_k^{op}}{\gamma_k} + 1\right)^{-1} - \lambda = 0 \text{, if } \tilde{f}_k^{op} > 0 \text{, } k = 1, 2, \ldots, K \tag{10}$$

$$\psi_k \; - \; \lambda < 0 \text{, if } \tilde{f}_k^{op} = 0 \text{, } k = 1, 2, \ldots, K.$$

The optimal fractional allocation satisfies the conditions in Equation (10) plus the constraint $\sum_{k=1}^{K} \tilde{f}_k^{op} = 1$. The unit sum constraint for the fractions implies that only K–1 of the $\tilde{f}_k^{op}$ values need to be estimated, since the fractional allocation to one good is automatically determined from the fractional allocations of all the other goods. To accommodate this constraint, designate activity purpose 1 as a purpose to which the individual allocates some non-zero fraction of consumption. For the first good, the KKT condition may then be written as:

$$\lambda = \psi_1 \left(\frac{\tilde{f}_1^{op}}{\gamma_1} + 1\right)^{-1} \tag{11}$$

Substituting for λ from above into Equation (8) for the other goods (k = 2,…, K), and taking logarithms, we can rewrite the KKT conditions as:

$$\tilde{\eta}_k = \varepsilon_k - \varepsilon_1 = \tilde{V}_k \; (= V_k - V_1) \quad \text{if fractional split} = \tilde{f}_k^{op} \; (k = 2, 3, \ldots, K), \; \tilde{f}_k^{op} > 0 \; (k = 2, 3, \ldots, K)$$

$$\tilde{\eta}_k = \varepsilon_k - \varepsilon_1 < \tilde{V}_{k0} \; (= V_{k0} - V_1) \text{ if } \tilde{f}_k^{op} = 0 \; (k = 2, 3, \ldots, K) \text{, where} \tag{12}$$

$$V_k = -\boldsymbol{\beta}'\mathbf{z}_k + \frac{1}{\sigma}\ln p_k + \ln\left(\frac{\tilde{f}_k^{op}}{\gamma_k} + 1\right) \; (k = 1, 2,, \ldots, K) \text{, and } V_{k0} = -\boldsymbol{\beta}'\mathbf{z}_k + \frac{1}{\sigma}\ln p_k \; (k = 1, 2, \ldots, K).$$

Then, we get the expression below in the reverse Gumbel MDCEV (or RG-MDCEV) model for the fractional allocation pattern, where the first M inside goods are consumed at levels $\tilde{f}_k^{op}$ ($k = 2, 3, \ldots, M$), and $\tilde{f}_1^{op} = 1 - \sum_{k=2}^{M} \tilde{f}_k^{op}$:

$$P\left(\tilde{f}_2^{op},...,\tilde{f}_M^{op},0,0,...,0\right)$$

$$=|J|\int_{\tilde{\eta}_{M+1}=-\infty}^{\tilde{\eta}_{M+1}=\tilde{V}_{M+1,0}}\int_{\tilde{\eta}_{M+2}=-\infty}^{\tilde{\eta}_{M+2}=\tilde{V}_{M+2,0}}\cdots\int_{\tilde{\eta}_K=-\infty}^{\tilde{\eta}_K=\tilde{V}_{K,0}}\boldsymbol{f}_{\eta}(\tilde{V}_2,\tilde{V}_3,...,\tilde{V}_M,\tilde{\eta}_{M+1},\tilde{\eta}_{M+1},\ldots,\tilde{\eta}_K)\,d\tilde{\eta}_{M+1}d\tilde{\eta}_{M+2},...,d\tilde{\eta}_K$$

$$=|J|\left.\frac{\partial^M \boldsymbol{F}_{\tilde{\boldsymbol{\eta}}}(\tilde{\eta}_2,\tilde{\eta}_3,...,\tilde{\eta}_M,\tilde{V}_{M+1,0},\tilde{V}_{M+2,0},\ldots,\tilde{V}_{K,0})}{\partial\tilde{\eta}_2\partial\tilde{\eta}_3...\partial\tilde{\eta}_M}\right|_{\tilde{\eta}_2=\tilde{V}_2,\tilde{\eta}_3=\tilde{V}_3,...,\tilde{\eta}_M=\tilde{V}_M}$$

$$=|J|(M-1)!\left[\frac{\exp\left(\sum_{i=1}^{M}V_k\right)}{\left(\sum_{k=1}^{M}\exp(V_k)\right)^M}+\sum_{D\subset\{M+1,M+2,...,K\},|D|\geq 1}(-1)^{|D|}\frac{\exp\left(\sum_{i=1}^{M}V_k\right)}{\left(\sum_{k=1}^{M}\exp(V_k)+\sum_{k\in D}\exp(V_{k0})\right)^M}\right] \qquad (13)$$

where $|J| = \left(\prod_{i=1}^{M} c_i\right)\left(\sum_{i=1}^{M}\frac{1}{c_i}\right)$, where $c_i=\left(\frac{1}{\tilde{f}_i^{op}+\gamma_i}\right)$.

The probability that all the inside goods are consumed at fractional levels $\tilde{f}_2^{op},\tilde{f}_3^{op},...,\tilde{f}_K^{op}$ is:

$$P\left(\tilde{f}_2^{op},\tilde{f}_3^{op},...,\tilde{f}_K^{op}\right)$$

$$=|J|\,\boldsymbol{f}_{\tilde{\boldsymbol{\eta}}}(V_2,V_3,...,V_K)=|J|(K-1)!\frac{\exp\left(\sum_{i=1}^{K}V_k\right)}{\left(\sum_{k=1}^{K}\exp(V_k)\right)^K} \qquad (14)$$

The probability that none of the inside goods are consumed is:

$$P(0,...,0)=1+\sum_{D\subset\{2,...,K\},|D|\geq 1}(-1)^{|D|}\frac{1}{\left(1+\sum_{k\in D}e^{\tilde{V}_{k0}}\right)}=1+\sum_{D\subset\{2,...,K\},|D|\geq 1}(-1)^{|D|}\frac{e^{V_{10}}}{\left(e^{V_{10}}+\sum_{k\in D}e^{V_{k0}}\right)}. \qquad (15)$$

The parameters to be estimated in the model above include the $\boldsymbol{\beta}$ vector, the $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,...\gamma_K)$ vector, and the σ scalar. However, note that these parameters from the RG-MDCEV fractional model also appear in the total budget model, and hence we discuss the overall estimation procedure in Section 2.3 after first discussing the total budget (in the product group) and the linking approach between the fractional allocation model and the total budget model.

2.2. Total Budget in Product Group and Linking with the RG-MDCEV Model

2.2.1. Theoretical Background

This section is motivated by the basic idea of two-stage budgeting based on separable direct utility functions across mutually disjunct product groups; that is, the overall utility from consuming in different product groups is the sum of separable sub-utilities for each product group. This implies that once the budget for any specific product group is known, the optimal allocation to specific commodities within the group is solely a function of the group budget and the vector of prices of individual commodities within the group. The two-stage budgeting concept is then based on the notion that, given an overall separable utility function across product groups, the total overall budget can first be split across group-level budgets and then followed by the allocation of the group budget over the individual goods within the product group category under consideration (see Strotz, 1957). Gorman (1959, 1961), building off the two-stage budgeting idea, examined the conditions under which the first level allocation of the total budget to different product groups could be undertaken without detailed information about the prices of individual goods within each product group. That is, the conditions under which a specific group of commodities can be treated as a single commodity with a single group price index so that, at the first level, the total budget across all product categories can be split up across groups based only on the single price index for each group. For this, each group indirect utility function should satisfy Gorman's polar form (GPF) for each product category.

2.2.2. Unconditional Demand

To set the framework, we first propose a logarithm form quality-adjusted price index for each inside good *k* within the product group *G* of interest as follows:

$$-\ln\tilde{\pi}_k = \frac{1}{1/\sigma}\left[\ln\left(\psi_k\gamma_k + \gamma_k\sum_{\substack{j=1\\ j\neq k}}^{M}\gamma_j\left(\psi_k - \psi_j\right)\right)\right], \text{ or } \tilde{\pi}_k^{-1/\sigma} = \psi_k\gamma_k + \gamma_k\sum_{\substack{j=1\\ j\neq k}}^{M}\gamma_j\left(\psi_k - \psi_j\right). \qquad (16)$$

For now, assume that $\tilde{\pi}_k^{-1/\sigma} > 0 \ \forall k$ (we will revisit this issue in the next section). The second term on the right side of the expression above is a summation over all consumed goods $(j-1,2,...,M)$. The reason for this specific form will become clear later. The negative sign in front of $\ln\tilde{\pi}_k$ on the left side of the above equation is because $\tilde{\pi}_k$ is a price index, which should increase or decrease as the actual price of good *k* (that is, p_k as embedded within ψ_k) increases or decreases. Similarly, as γ_k increases, satiation for good *k* decreases and the desirability of the good increases, which is similar to the effect of a price decrease. In effect, then, in logarithm form, $-\tilde{\pi}_k$ represents the total quality-adjusted utility of good *k*, which is then normalized by the coefficient on the actual log(price) variable (representing the marginal utility of income), so that $\ln\tilde{\pi}_k$ represents, in log form, the quality-adjusted price index. Such a logarithmic form has been used for single discrete choice models in Rouwendal and Boter, 2009 and Truong and Hensher, 2014, though our formulation above is for a multiple discrete fractional allocation model; also of note is that $\tilde{\pi}_k^{-1/\sigma} = (1/\tilde{\pi}_k^{1/\sigma})$ is another (reciprocal-based) form of a quality-adjusted utility of good *k*, which, like $-\tilde{\pi}_k$, increases as the price p_k decreases and decreases as the price p_k increases (we will revisit this issue again later for a more crisp interpretation of $\tilde{\pi}_k^{-1/\sigma}$).

Collect the quality-adjusted price indices $\tilde{\pi}_k$ for the inside goods of product category *G* in a vector $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \tilde{\pi}_2, ..., \tilde{\pi}_K)$. Consider the following indirect utility function form for group *G*, assuming, as discussed earlier, that the product category *G* is separable from other groups of commodities within the overall direct utility function across all product groups:

$$W(y^*, \tilde{\boldsymbol{\pi}}) = \frac{y^*}{b(\tilde{\boldsymbol{\pi}})} + a(\tilde{\boldsymbol{\pi}}) = q^* + a(\tilde{\boldsymbol{\pi}}), \tag{17}$$

where y^* is the budget allocated to group *G*, $b(\tilde{\boldsymbol{\pi}})$ is a group price index for group *G* (which is a composite group price index used at the first stage in the two-stage budgeting that is a function of the quality-adjusted price indices of the inside goods; that is, the elements of the vector $\tilde{\boldsymbol{\pi}}$), $a(\tilde{\boldsymbol{\pi}}))$ is another function of $\tilde{\boldsymbol{\pi}}$, and q^* is the total quantity consumed across all elementary commodities within product group *G*. To be able to use a single group price index for Group *G* (essentially viewing the entire product group *G* as one single commodity with a single price index of $b(\tilde{\boldsymbol{\pi}})$, so that $q^* = \frac{y^*}{b(\tilde{\boldsymbol{\pi}})}$), Equation (17) needs to be consistent with GPF. That is, $b(.)$ needs to be homogenous of degree 1 and $a(\tilde{\boldsymbol{\pi}})$ homogenous of degree 0. As in Hausman et al. (1995), we assume that $a(\tilde{\boldsymbol{\pi}})$ is a constant. So, we use the constant elasticity of substitution (CES) formulation and write:

$$b(\tilde{\boldsymbol{\pi}}) = \left[\sum_k \left(\left(\tilde{\pi}_k \right)^{-1/\sigma} \right) \right]^{-\sigma}, \tag{18}$$

which is clearly homogenous of degree 1 as required because

$$\left[\sum_k \left(\alpha \tilde{\pi}_k \right)^{-1/\sigma} \right]^{-\sigma} = \alpha b(\tilde{\boldsymbol{\pi}}). \tag{19}$$

Also, as desired, the group price index $b(\tilde{\boldsymbol{\pi}})$ increases as the price p_k of any good *k* increases (note that $\sigma > 0$).

To be sure, the GPF is essentially the condition that allows the consideration of an entire set of elementary commodities in a product group to be represented as one aggregate commodity with a singular price. In the specific context of the current discussion, defining q_k^* as the quantity allocated to good *k* within the product category *G* (so, $q^* = \sum_{k=1}^{K} q_k^*$), the GPF conditions on $b(\tilde{\boldsymbol{\pi}})$ and $a(\tilde{\boldsymbol{\pi}})$ in Equation (17) is the vehicle that allows us to view the entire product group *G* as one aggregate commodity with a composite price index of $b(\tilde{\boldsymbol{\pi}})$. To see this, applying Roy's identity to the Gorman polar indirect utility form of Equation (17) with $a(\tilde{\boldsymbol{\pi}})$ as a constant, one can write the Marshallian demand for good *k* within the product group *G* as follows:

$$q_k^* = -\frac{\partial W(y^*,\tilde{\boldsymbol{\pi}})/\partial\tilde{\pi}_k}{\partial W(y^*,\tilde{\boldsymbol{\pi}})/\partial y^*} = y^*\left(\frac{1}{b(\tilde{\boldsymbol{\pi}})}\right)\left(\frac{\partial b(\tilde{\boldsymbol{\pi}})}{\partial\tilde{\pi}_k}\right) = \frac{1}{\tilde{\pi}_k}\left(\frac{\tilde{\pi}_k^{-1/\sigma}}{\sum_k \tilde{\pi}_k^{-1/\sigma}}\right)y^*,\ k = 1,2,...,K. \qquad (20)$$

The above equation implies that $y^* = b(\tilde{\pi})q^* = b(\tilde{\pi})\left(\sum_{k=1}^{K} q_k^*\right) = \sum_{k=1}^{K}\tilde{\pi}_k q_k^*.$ Also, note that another consequence of the GPF condition is that the total budget y^* allocated to product group *G* is only dependent on the total consumption q^* in the product group, and entirely independent of the allocation (distribution) of the budget y^* to individual elementary commodities within the product group. The result is that, as in Hausman et al. (1995), the total budget cannot appear as a determinant of the fractional split MDCEV model (in other words, the GPF requires that the group price index-based linking function $b(\tilde{\boldsymbol{\pi}})$ have only the quality adjusted price index vector $\tilde{\boldsymbol{\pi}}$ as an argument, and cannot include y^*). Further, the linking function needs to be exogenous and strictly upward from the second stage fractional split model to the first stage total budget model (this has implications in the empirical specification).

For further reference before moving forward, note that, corresponding to the demand of Equation (20), we can also write the fractional demand for (allocation of the budget y^* to) good *k* as follows:

$$\frac{q_k^*\tilde{\pi}_k}{y^*} = \tilde{f}_k^* = \left(\frac{\tilde{\pi}_k^{-1/\sigma}}{\sum_k \tilde{\pi}_k^{-1/\sigma}}\right),\ k = 1,2,...,K. \qquad (21)$$

2.2.3. Conditional Demand

Equations (20) and (21) provide the notional demand and notional fractional demand, respectively, on good *k*, because they are critically predicated on the assumption that $\tilde{\pi}_k^{-1/\sigma} > 0$ for all inside goods *k*. In reality, though, there is nothing to prevent $\tilde{\pi}_k^{-1/\sigma}$ from being zero or even negative. As we will now show, the condition $\tilde{\pi}_k^{-1/\sigma} > 0\ \forall\ k$ is equivalent to assuming that all the inside goods are chosen for consumption. That is, the indirect utility of Equation (17) corresponds to a direct utility maximization problem that does not include the non-negativity constraints on the inside good consumptions and assumes that $y^* \geq 0$. So, the entire discussion in the previous section assumes that these non-negativity constraints are automatically honored. But let's consider the more realistic case when, conditional on consumption in the product category (that is, still assuming $y^* > 0$), only goods 1 to *M* are consumed and others are not. In this situation, following Lee and Pitt (1986), the notional fractional demands can be viewed as latent variables that correspond to the actual observed optimal fractional consumptions $\tilde{f}_k^{op}$ in a specific way, which is based on recognizing the non-negativity constraints. Use a transformation from the vector $\tilde{\boldsymbol{\pi}}$ to a strictly non-negative vector of virtual prices $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_K)$ that support the non-negative fractional demands, conditional on $y^* > 0$, such that $\pi_k^{-1/\sigma} = \tilde{\pi}_k^{-1/\sigma}$ if $\tilde{\pi}_k^{-1/\sigma} \geq 0$; $\pi_k^{-1/\sigma} = 0$ if $\tilde{\pi}_k^{-1/\sigma} < 0$. As we show in Appendix B, the inequality $\tilde{\pi}_k^{-1/\sigma} < 0$, in fact, is just another way of writing the condition for zero consumption of good *k* based on the KKT conditions in Equation (10) (that is, the strictly non-negative virtual price vector $\boldsymbol{\pi}$ represents shadow prices corresponding to the indirect utility function of Equation (17), but now accommodating the non-negativity constraints of the direct utility function). With this, we can define a switching regime where good *k* is consumed if $\tilde{\pi}_k^{-1/\sigma} \geq 0$ $(\pi_k^{-1/\sigma} = \tilde{\pi}_k^{-1/\sigma})$ and not consumed if $\tilde{\pi}_k^{-1/\sigma} < 0$ $(\pi_k^{-1/\sigma} = 0)$. In this regard, $\tilde{\pi}_k^{-1/\sigma}$ may be viewed as the threshold quality-adjusted utility of good *k* that needs to be positive for positive consumption of good *k*. Then, the non-negative observed fractional demands corresponding to Equation (21) may be written as follows:

$$\tilde{f}_k^{op} = \left(\frac{\pi_k^{-1/\sigma}}{\sum_k \pi_k^{-1/\sigma}} \right) \Bigg| y^* > 0, \; k = 1, 2, ..., K. \tag{22}$$

Thus, from an economic standpoint, the second stage expenditure-based conditional (on consumption in the product group) fractional allocation MDCEV model should be:

$$\tilde{f}_k^{op} = \left(\frac{\pi_k^{-1/\sigma}}{\sum_{k=1}^{K} \pi_k^{-1/\sigma}} \right) = \left(\frac{\tilde{\pi}_k^{-1/\sigma}}{\sum_{m=1}^{M} \tilde{\pi}_m^{-1/\sigma}} \right) = \frac{\psi_k \gamma_k + \gamma_k \sum_{\substack{j=1 \\ j \neq k}}^{M} \gamma_j \left(\psi_k - \psi_j \right)}{\sum_{j=1}^{M} \psi_j \gamma_j}, \; k = 1, 2, ..., M \text{ (for consumed goods)} \tag{23}$$

$$\tilde{f}_k^{op} = 0, \; k = M+1, M+2, ..., K \text{ (for non-consumed goods)}$$

The denominator in the first expression above takes the form as shown because of the following equality:

$$\sum_{m=1}^{M} \tilde{\pi}_m^{-1/\sigma} = \sum_{m=1}^{M} \left(\psi_m \gamma_m + \gamma_m \sum_{\substack{j=1 \\ j \neq m}}^{M} \gamma_j \left(\psi_m - \psi_j \right) \right) = \sum_{m=1}^{M} \psi_m \gamma_m, \text{ because } \sum_{m=1}^{M} \left(\gamma_m \sum_{\substack{j=1 \\ j \neq m}}^{M} \gamma_j \left(\psi_m - \psi_j \right) \right) = 0. \tag{24}$$

Thus, from a theoretical standpoint, using the unconditional linking form of Equation (18) as the product group price index from the expenditure-based fractional split model to the total expenditure (in product group) first stage model is consistent with utility-based two-stage budgeting as long as the (conditional on product category consumption) fractional split model among the consumed inside goods takes the form in Equation (23). As we discuss later, Equation (23) is exactly the form for forecasting the fractional splits among consumed goods in the product category, conditional on positive investment in the product category. Thus, our proposed theoretical framework with the unconditional linking form of Equation (18) is consistent with two-stage utility-based economic theory. Also, the MDCEV fractional allocation model is to be interpreted as a conditional budget share model, and not a quantity share model, consistent with the discussion earlier in this section.

Finally, the entire analysis above is predicated on $y^* > 0$. That is, fractional consumptions of inside goods within the product category *G* are relevant only if $y^* > 0$. So, we treat y^* as a latent variable for modeling purposes, and relate the actual observed budget y to y^* such that $y = y^*$ if $y^* > 0$; $y = 0$ if $y^* \leq 0$. This positivity condition is based on the structure of first budget equation, as discussed next.

2.2.4. Total Budget (in Product Group) Empirical Specification

In the empirical specification, we use the linking function $b(\tilde{\boldsymbol{\pi}}) = \left[\sum_k \left(\left(\tilde{\pi}_k\right)^{-1/\sigma} \right)\right]^{-\sigma} = \left[\sum_k \psi_k \gamma_k\right]^{-\sigma}$. As discussed in Section 2.2.2, the linking function has to be completely exogenous to the total budget equation. This generates problems if the stochasticity in $\psi_k = \mu_k \exp\left(\varepsilon_k\right)$ is carried over as such into the total budget first stage equation, because this engenders a correlation between the first stage budget equation and the second stage MDCEV model. This can be addressed in one of two ways. The first is to use the expected value of $b(\tilde{\boldsymbol{\pi}}) = \left[\sum_k \psi_k \gamma_k\right]^{-\sigma}$ (or the expected value of the logarithm of $b(\tilde{\boldsymbol{\pi}})$) as the deterministic linking function. The second is to use a different set of error terms τ_k instead of ε_k in the linking function (with τ_k being independent of ε_k), so that $\psi_k = \mu_k \exp\left(\tau_k\right)$. In this paper, we adopt the second approach, which, unlike the first approach, recognizes unobserved individual heterogeneity in the linking effect. We assume for convenience that the error terms τ_k are independent across inside goods, The linking function then may be written as:

$$b(\tilde{\boldsymbol{\pi}}) = \left[\sum_k \left(\mu_k \exp\left(\tau_k\right)\right)\gamma_k\right]^{-\sigma} = \left[\sum_k a_k \exp\left(\tau_k\right)\right]^{-\sigma}, \text{ where } a_k = \mu_k \gamma_k. \tag{25}$$

The linking function $b(\tilde{\boldsymbol{\pi}})$ is always positive, by construction. The total budget y needs to be non-negative, though it can take the value of zero. For modeling purposes, consider the censored Tobit regression equation:

$$\begin{aligned} y^* &= \boldsymbol{\theta}'\mathbf{s} - \lambda \ln b(\tilde{\boldsymbol{\pi}}) - \lambda\zeta = \boldsymbol{\theta}'\mathbf{s} + \lambda\sigma\left[\ln \sum_k a_k \exp(\tau_k)\right] - \lambda\zeta \\ &= \boldsymbol{\theta}'\mathbf{s} - \lambda\eta, \text{ with } \eta = \left(\zeta - \sigma\left[\ln \sum_k a_k \exp(\tau_k)\right]\right) \end{aligned} \tag{26}$$

$$y = \begin{cases} 0 & \text{if } y^* \leq 0 \\ y^* & \text{if } y^* > 0 \end{cases}$$

where $\mathbf{s}$ is an exogenous variable vector, $\boldsymbol{\theta}$ is a corresponding coefficient vector, λ is a scalar link parameter (λ>0), and ζ is a random variable capturing the effects of unobserved variables. The linking parameter appendage to the error term ζ in the first line of Equation (26) is innocuous, and is only for presentation ease in the characterization of the error term η. As the price of any inside good *k* (p_k) decreases, or as the non-cost systematic (log) baseline utility element for any inside good *k* ($\boldsymbol{\beta}'\mathbf{z}_k$) increases, the value of η falls, and the value of the budget allocated to group *G*, y, increases.

To continue with the model formulation, one needs the distribution of the underlying latent variable y^*, or, equivalently, the distribution of the random variable η. We now assume that the random variable ζ is reverse Gumbel distributed with a scale σ. With this assumption, the distribution of the random variable η takes a surprisingly elegant form because the survival distribution function (SDF) of the difference between a reverse Gumbel distribution with scale σ and σ times the logarithm of the weighted sum of independent standard exponentially distributed random variables has a closed form (note that $\exp(\tau_k)$ is standard exponentially distributed, because τ_k is standard reverse-Gumbel; also, note that, by construction, a_k>0 for all *k*). This univariate distribution for η, to our knowledge, has not appeared in the statistical literature, but is what we will refer to as the minLogistic distribution.[7] The precise distributional shape of η will depend on the values of a_k, but the distribution is skewed toward the left, similar to that of a reverse Gumbel distribution (except for the case when *K*=1, in which case the minLogistic distribution collapses to a simple symmetric logistic distribution).[8,9] Figure 1

[7] This distribution was proposed by Bhat recently and applied in a pair of recent papers for a different application; see Mondal and Bhat, 2021 and Bhat et al., 2022; but much more of its properties are derived and formally presented here; these properties will be put to good use not only in estimation, but also in forecasting, as discussed later.

[8] Note also that, because η appears in negative form on the right side of Equation (26), the effective distribution for y^*, takes a distributional form that has a long right tail. This is consistent with expenditure data or mileage data, which is rarely symmetric, and has a long trailing right edge. This is also the reason why a lognormal form has been used to model such expenditure data. In our case here, the convenience of the distribution form with linking leads us to use the reverse Gumbel distribution for ζ, which translates to the right skew of y^*. Interestingly, when there is no linkage, the only error term appearing in Equation (26) is the reverse Gumbel variable ζ, which still lends the desirable right-skew to the distribution for y^*.

[9] When there is no linkage, $\lambda\sigma$ serves as the scale parameter of the error term in the Tobit regression of Equation (26) (λ and σ are identified separately only if there is price variations across the goods in the product category under consideration in the fractional MDCEV model; otherwise, σ needs to be normalized to one and λ is identified). With linkage, the linkage parameter λ (that captures the systematic baseline preference effects as well as unobserved heterogeneity of the baseline preference effects on the overall price index for the product category) is estimable, but σ again is estimable only if there is price variation across commodities in the product group. However, with linkage, heteroscedasticity is also introduced in the Tobit model because of the τ_k error terms, as will be noted later in Equation (30). Finally, if one uses the empirical specification where the expected value of

provides a sample plot of the distribution for a situation with three goods (K=3) with $a_1 = 1$, $a_2 = 2$, and $a_3 = 3$, and $\sigma = 1$. Additional properties of this new minLogistic distribution of the random variable η $(a_k > 0\ \forall\, k)$ with scale σ are now presented (with proofs) below (the expressions have also been verified through simulation experiments):

Property 1

With the above-mentioned distributional assumptions on the error terms ε_k and ζ, the survival distribution function (SDF) of η takes a closed-form as follows (see Appendix A.1 for the derivation):

$$S_\eta(t) = \text{Prob}(\eta > t) = \frac{1}{\prod_{k=1}^{K}\left(1 + a_k e^{t/\sigma}\right)},\ a_k > 0\ \forall\, k. \tag{27}$$

The corresponding cumulative distribution function and density functions are readily obtained as[10]:

$$F_\eta(t) = \text{Prob}(\eta < t) = 1 - \left[\frac{1}{\prod_{k=1}^{K}\left(1 + a_k e^{t/\sigma}\right)}\right],\ a_k > 0\ \forall\, k,$$
$$f_\eta(t) = \frac{1}{\sigma} e^{t/\sigma} \times S_\eta(t) \times \sum_k \left(\frac{a_k}{1 + a_k e^{t/\sigma}}\right),\ a_k > 0\ \forall\, k. \tag{28}$$

Property 2

The minLogistic distribution is strongly unimodal (see Appendix A.2). The mode does not have a closed-form expression in the general case when K>1, but is the solution to the following equation that can be obtained numerically (see Appendix A.2):

$b_G(\boldsymbol{\pi}) = \left[\sum_k \psi_k \gamma_k\right]^{-\sigma}$ (or its monotonic transformation, such as a logarithm transformation) is employed as the deterministic linking function, it is possible to separately identify the linking parameter λ from a separate scale parameter for the ζ error term as well as σ in the case of price variation in the fractional MDCEV model. If there is no price variation, σ has to be normalized to one, but the linking parameter λ again is identifiable separately from the scale parameter for the ζ error term. Essentially, this alternative deterministic linking specification, which we do not consider in this paper, adds a parameter while maintaining homoscedasticity in the Tobit model. The deterministic linking specification also nests the unlinked model. Our empirical stochastic linking approach, on the other hand, has the same number of parameters as the unlinked model, while also engendering heteroscedasticity in the Tobit equation (see also Section 4.4.2). Thus, it is parsimonious and affords more flexibility, though the error term distribution of the Tobit model gets altered; so our empirical model does not nest the unlinked model.

[10] When K=1, it is straightforward to note that the distribution of η takes a symmetric Logistic distribution, with a location parameter of $-\ln a_k$ and a scale parameter of σ.

$$\varpi = Mode(\eta) \Rightarrow e^{\varpi/\sigma}\left[\sum_{k=1}^{K}\xi_k(\varpi) + \frac{\sum_{k=1}^{K}\xi_k^2(\varpi)}{\sum_{k=1}^{K}\xi_k(\varpi)}\right] = 1, \text{ where } \xi_k(\varpi) = \frac{a_k}{1+a_k e^{\varpi/\sigma}}. \tag{29}$$

For the case when K=1, the minLogistic distribution collapses to a symmetric logistic distribution, and there is a closed form solution for the mode, which occurs at $\varpi = (1/a)\times\sigma$, as can be readily seen by applying the formula above for K=1. For the case of the minLogistic distribution plotted in Figure 1, application of the Equation (29) reveals a mode at –1.78, which, as can be seen from Figure 1, is the point at which the density function reaches its peak value.

Property 3

The mean and variance of the minLogistic distribution are given by the following closed-form expressions (see Appendix A.3):

$$E(\eta) = -\sigma\left[\sum_{k=1}^{K}\frac{a_k^{K-1}\ln(a_k)}{\prod_{\substack{j=1\\ j\neq k}}^{K}\left(a_k - a_j\right)}\right] \text{ and}$$

$$Var(\eta) = E(\eta^2) - \left[E(\eta)\right]^2 = \sigma^2\left[\sum_{k=1}^{K}\left(\frac{3a_k^{K-1}\ln^2(a_k) + \pi^2 a_k^{K-1}}{3\times\prod_{\substack{j=1\\ j\neq k}}^{K}\left(a_k - a_j\right)}\right)\right] - \left[E(\eta)\right]^2 \tag{30}$$

In case one or more of the a_k values are equal, the corresponding expressions become rather unwieldy to write in a generic form; however, the nice and elegant expressions above will provide almost the exact values if applied after introducing very small perturbations to the a_k values to make them unequal. Note also that, in our current application, the a_k values will generally not be the same because of alternative specific parameters in the baseline preference utilities. An important issue to note above is that the variance of the error term η is heteroscedastic across individuals (as a_k would vary across individuals), an issue we will get back to later in this paper.[11]

[11] An important derivative of the results in Equation (30) is that, to the author's knowledge, this is the first time in the statistical literature that the expected value and variance of the random variable $\xi = \sigma\left[\ln\sum_k a_k \exp(\tau_k)\right]$ has also been derived. Specifically, $E(\xi) = -\vec{\gamma}\sigma - E(\eta)$, and $Var(\xi) = Var(\eta) - \frac{\vec{\pi}^2\sigma^2}{6}$, where $\vec{\gamma}$ is the Euler's constant and $\vec{\pi}$ is the usual ratio of the circumference of a circle to its diameter.

Property 4

The mean and variance of the minLogistic distribution truncated from above at the point c are given by the following simple expressions (see Appendix A.4):

$$E(\eta)\big|(\eta < c) = c - \sigma\left(\frac{1}{F_\eta(c)}\right)\left[\sum_{k=1}^{K}\left(\frac{a_k^{K-1}\ln(1+a_k e^{c/\sigma})}{\prod\limits_{\substack{j=1\\ j\neq k}}^{K}\left(a_k - a_j\right)}\right)\right] \text{ and}$$

$$Var(\eta)\big|(\eta < c) = \left(\sigma^2\left(\frac{1}{F_\eta(c)}\right)\left[-2\times\sum_{k=1}^{K}\left(\frac{a_k^{K-1}Li_2(-a_k e^{c/\sigma})}{\prod\limits_{\substack{j=1\\ j\neq k}}^{K}\left(a_k - a_j\right)}\right)\right]\right) - \left[c - E(\eta)\big|(\eta < c)\right]^2, \tag{31}$$

where $Li_2(h) = -\int\limits_{t=0}^{h}\frac{\ln(1-t)}{t}dt$ in the variance expression represents the dilogarithm function and is easily computed using one-dimensional integration (even if not having an analytical expression). However, note that the expected value of the minLogistic distribution truncated from above (first expression in Equation (31)) is a closed-form expression.

With the above properties in hand, and defining $\vartheta = \lambda\sigma$, the cumulative distribution of y^* may be derived from Equation (26) as:

$$F_{y^*}(t) = \text{Prob}(y^* < t) = \text{Prob}(\boldsymbol{\theta}'\mathbf{s} - \lambda\eta < t) = \text{Prob}\left[\eta > \left(\frac{\boldsymbol{\theta}'\mathbf{s} - t}{\lambda}\right)\right] = \frac{1}{\prod\limits_{k=1}^{K}\left(1 + a_k e^{(\boldsymbol{\theta}'\mathbf{s}-t)/\vartheta}\right)}, \tag{32}$$

and the corresponding density function is:

$$f_{y^*}(t) = \frac{1}{\vartheta}\left(e^{[(\boldsymbol{\theta}'\mathbf{s}-t)/\vartheta]}\right)\times F_{y^*}(t)\times\sum_k\frac{a_k}{\left(1 + a_k e^{[(\boldsymbol{\theta}'\mathbf{s}-t)/\vartheta]}\right)}. \tag{33}$$

3. MODEL ESTIMATION AND FORECASTING

Collect the parameters to be estimated in a vector $\tilde{\boldsymbol{\mu}} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \sigma, \boldsymbol{\theta}, \lambda)'$. The likelihood function corresponding to no allocation to product group G is given by:

$$L(\tilde{\boldsymbol{\mu}})\text{=Prob}(y^* < 0) = F_{y^*}(0) = \frac{1}{\prod\limits_{k=1}^{K}\left(1 + a_k e^{(\boldsymbol{\theta}'\mathbf{s})/\vartheta}\right)}. \tag{34}$$

The likelihood corresponding to non-zero allocation to the product group with a budget allocation of h , and the fractional allocation to the first M inside goods in the product group is:

$$L(\boldsymbol{\mu}) = f_{y^*}(h) \times P\left(\tilde{f}_2^*, ..., \tilde{f}_M^*, 0, 0, ..., 0\right), \tag{35}$$

where the probability of the fractional allocation on the right side of the above equation is provided by Equation (13). Similar functions may be written for the case of a budget allocation of y and fractional allocation to all inside goods (based on Equation (14)) and the case of a budget of y and fractional allocation to only one inside good (based on Equation (15)).

The above model formulation does not consider unobserved heterogeneity in the sensitivity to exogenous variables. This may be introduced in a straightforward way by allowing the $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ parameters to be randomly distributed. For example, assuming a specific continuous parametric distribution (say $\vec{f}$) for $\boldsymbol{\psi} = (\boldsymbol{\beta}', \boldsymbol{\theta}')'$ with an underlying parameter vector $\tilde{\boldsymbol{\psi}}$, the likelihood function for the case of a budget allocation of y, and the fractional allocation to the first M inside goods in the product group, is:

$$L(\tilde{\boldsymbol{\psi}}, \boldsymbol{\gamma}, \sigma, \vartheta) = \int\limits_{\boldsymbol{\psi}} f_{y^*}(h) \times \text{Prob}[\tilde{f}_2^*, ..., \tilde{f}_M^*, 0, 0..., 0] \, \vec{f}(\boldsymbol{\psi}; \tilde{\boldsymbol{\psi}}) \, \mathbf{d}\boldsymbol{\psi}. \tag{36}$$

In the random parameter specification above, there should be no covariance across the elements of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ to preserve strict exogeneity of the linking function in the first stage budgeting equation. The log-likelihood function may be developed across all individuals, and the parameters may be estimated using maximum likelihood estimation-based approaches.

For model forecasting, it is easy enough to predict the total expected budget allocated to the entire product category using the properties of the minLogistic distribution as follows:

$$\begin{aligned}
E(y) &= 0 \times P(y^* < 0) + \left(E(y^*) \mid y^* > 0\right) \times P(y^* > 0) = \left(E(y^*) \mid y^* > 0\right) \times P(y^* > 0) \\
&= \left[E(\boldsymbol{\theta}'\mathbf{s} - \lambda\eta) \mid \left(\boldsymbol{\theta}'\mathbf{s} - \lambda\eta > 0\right)\right] \times P(\boldsymbol{\theta}'\mathbf{s} - \lambda\eta > 0) \\
&= \left[\boldsymbol{\theta}'\mathbf{s} - \lambda E(\eta) \mid \left(\eta < \frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda}\right)\right] \times P\left(\eta < \frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda}\right). \\
&= \left[\boldsymbol{\theta}'\mathbf{s} - \lambda \left(\frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda} - \sigma \left(\frac{1}{F_\eta\left(\frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda}\right)}\right) \left[\sum_{k=1}^{K} \left(\frac{a_k^{K-1} \ln\left(1 + a_k e^{\frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda\sigma}}\right)}{\prod\limits_{\substack{j=1 \\ j \neq k}}^{K} \left(a_k - a_j\right)}\right)\right]\right]\right)\right] \times F_\eta\left(\frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda}\right) \text{ (Using Eq (31))} \\
&= \vartheta \left[\sum_{k=1}^{K} \left(\frac{a_k^{K-1} \ln\left(1 + a_k e^{\frac{\boldsymbol{\theta}'\mathbf{s}}{\vartheta}}\right)}{\prod\limits_{\substack{j=1 \\ j \neq k}}^{K} \left(a_k - a_j\right)}\right)\right], \;\; \vartheta = \lambda\sigma \text{ and } a_k = \exp(\boldsymbol{\beta}'\mathbf{z}_k)\left(\frac{1}{p_k^{1/\sigma}}\right)
\end{aligned} \tag{37}$$

The expression above may be used to forecast the expected value of the total budget allocation for the product category for any individual, and also may be used to compute elasticity effects of variables that directly impact the overall budget allocated to the product group under consideration (that is, the effects of variables in the $\mathbf{s}$ vector) or that indirectly impact the overall budget allocated to the product group (that is, the effects of variables in the $\mathbf{z}_k$ vector or the price variable p_k on y). The variance maybe computed as follows, using the properties discussed earlier:

$$Var(y) = \left[Var(\boldsymbol{\theta}'\mathbf{s} - \lambda\eta) \,|\, \left(\eta < \frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda} \right) \right] \times P\left(\eta < \frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda} \right).$$

$$= \left(\vartheta^2 \left[2 \times \sum_{k=1}^{K} \left(\frac{a_k^{K-1} Li_2(-a_k e^{c/\sigma})}{\prod_{\substack{j=1 \\ j \neq k}}^{K} \left(a_k - a_j \right)} \right) \right] \right) - \left[E(y^*) \middle| \left(y^* > 0 \right) \right]^2 \tag{38}$$

An issue with applying the above formulas directly, though, is that the procedure would always result in some positive allocation to the product group. This may be okay when computing aggregate budget allocations to the product group across a set of individuals, but may not be appropriate when the forecasts are at an individual level to be embedded within an agent-based model with additional downstream models. An approach to preserve the possibility of zero allocation during forecasting is to first forecast the discrete allocation between zero units for the inside goods and some positive quantity for the goods. And then for those observations that are forecasted to have a positive allocation, the closed form expectation formula derived in this paper can be applied. The approach is as follows:

- Step 1: Compute the probability of a zero allocation for the product group *G* using Equation (34). Draw a random variable from the uniform distribution. If this draw is higher than the computed probability of zero allocation, declare a zero allocation for product group *G* for the observation and put $\hat{y}_k = 0 \ \forall \ k \ (k = 1, 2, ..., M)$. STOP.
- Step 2: The expected value for an observation with a non-zero allocation to the product category may be obtained as follows (again using Equation (31)):

$$E(y \,|\, y > 0) = \left[E(\boldsymbol{\theta}'\mathbf{s} - \lambda\eta) \,|\, \left(\boldsymbol{\theta}'\mathbf{s} - \lambda\eta > 0 \right) \right] = \vartheta \left(\frac{1}{F_\eta\left(\frac{\boldsymbol{\theta}'\mathbf{s}}{\lambda} \right)} \right) \left[\sum_{k=1}^{K} \left(\frac{a_k^{K-1} \ln\left(1 + a_k e^{\frac{\boldsymbol{\theta}'\mathbf{s}}{\vartheta}} \right)}{\prod_{\substack{j=1 \\ j \neq k}}^{K} \left(a_k - a_j \right)} \right) \right]. \tag{39}$$

[12]

[12] An important benefit of the proposed closed-form model in forecasting is that the first and second moments of the continuous consumption values are easily computed because of their closed form nature (as shown here in this equation for the first moment of the budget for the product category). This is unlike the case of the linear outside

However, the central purpose of a multiple discrete-continuous model is to predict the intensity of consumption (including potentially zero consumptions) of the inside goods. To do so, one needs to adopt a simulation technique to consistently predict the second stage fractional allocation among the goods in the product group as well as the intensity of total budget allocated to the entire product group. The procedure for predicting the MDCEV fractional allocation and actual amount of consumption to the inside goods (including possibly zero allocation) may be described as follows (continuing on from Step 2 from earlier), assuming a preset number of R draws of the error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_K)'$ to compute the expected value for each observation:

- Step 3: If procedure did not stop at Step 1 above as discussed there, set $r = 1$.
- Step 4: Draw K independent realizations of ε_k, one for each good k $(k = 1, 2, ..., K)$ from the reverse extreme value distribution with location parameter of 0 and the scale parameter equal to one; label this distribution as $\text{REV}(0,1)$. Compute ψ_k using Equation (3).
- Step 5: Re-order the goods in descending order of ψ_k; set a new index m ($m = 1, 2, ..., K$) for this new ordering of the inside goods. Let $\tilde{\psi}_m$ be the re-ordered vector of values of ψ_k and $\tilde{\gamma}_m$ be the re-ordered vector of values of γ_m.
- Step 6: Set $M = 2$ and $\tilde{f}_1 = 1$.
- Step 7: If $\tilde{\psi}_M < \dfrac{\tilde{\psi}_1}{\left(\dfrac{\tilde{f}_1}{\tilde{\gamma}_1} + 1\right)}$, set $\tilde{f}_m^{op} = 0$ $(m = M, M+1, ..., K)$ and go to Step 9. Else, set

$$\tilde{f}_m^{op} = \frac{\tilde{\psi}_m \tilde{\gamma}_m + \tilde{\gamma}_m \sum_{\substack{j=1 \\ j \neq m}}^{M} \tilde{\gamma}_j \left(\tilde{\psi}_m - \tilde{\psi}_j \right)}{\sum_{j=1}^{M} \tilde{\psi}_j \tilde{\gamma}_j}, \; m = 1, ..., M.$$ [13]

(40)

- Step 8: Set $M = M + 1$. If $M = K$, go to Step 9 (discussed later). Otherwise, go to Step 7.
- Step 9: If the intent is to predict consumption intensities for each of the inside goods across multiple individuals, compute the predicted expenditure intensity for each of the inside goods

good closed-form MDCEV utility models of Bhat (2018), Bhat et al. (2020) and Saxena et al. (2022a) in which forecasting becomes challenging because the first and second moments of optimal consumption of inside goods might not always be finite without an upper bound on the budget for the product category. Ironically, though, the situation of an infinite budget is precisely when the linear outside good utility models are technically applicable.

[13] The derivation of this expression is provided in Appendix B. It is straightforward to see that the fractional allocations among the consumed goods will sum to one, because the second term in the numerator of the expression above cancels out across the different consumed goods. As we show in Appendix B, the expression also guarantees that the predictions $\tilde{f}_m$ for any consumed good will be positive and less than 1. Note also that the form of this expression matches that needed for consistency with two-stage budgeting, as discussed in Section 2.2.3. Finally, the second term in the numerator of this expression is critical because it guarantees that the KKT condition of equal marginal utility at the point of actual expenditure on the consumed goods holds, as also discussed in Appendix B.

for the specific draw of the vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_K)'$ for each individual as the product of the predicted continuous allocation for the product group and the fractional budget allocation to each inside good: $\hat{y}_k = \tilde{f}_k^{op} \times E(y) \;\forall\; k \;(k = 1, 2, ..., K)$. But if the intent is to predict consumptions at an individual level for embedding into a larger prediction system, keep $\tilde{f}_k^{op}$ and compute $E(y \mid y > 0)$ per Step 2.

- Step 10: Set $r = r + 1$; if $r = R$, go to Step 11; otherwise go to Step 4.
- Step 11: If the goal is aggregate predictions, compute the mean of $\hat{y}_k$ (as computed from Step 9) for each good k across the different realizations, and declare that as the expenditure estimate. If the intent is to get predictions at an individual level, obtain the probability of $\tilde{f}_k^{op} = 0$ as the number of times out of R repetitions that the state occurred, and translate that probability into a deterministic assignment similar to Step 1 undertaken for the total budget allocation to the product group. If the fractional deterministic assignment is non-zero for good k, predict the budget allocated to good k for the individual as the average of the product of the non-zero $\tilde{f}_k^{op}$ values and the corresponding $E(y \mid y > 0)$ values across the R repetitions.

4. AN EMPIRICAL DEMONSTRATION

4.1. Background

Travel in the U.S. is predominantly undertaken using private motorized four-wheeler automobiles. This automobile dependence to pursue out-of-home activities may be traced to a number of reasons, including separation of residential locations from activity centers, relatively inexpensive costs of motorized vehicle ownership and maintenance, low gas prices, inadequate options to reach destinations by non-automobile means, and a traditional culture of valuing personal privacy and convenience. This dependence also has far-reaching impacts at multiple societal levels, including activity accessibility inequities across population segments at the individual/household level, and elevated levels of traffic congestion at the regional level (with ensuing repercussions on on-road vehicular travel determinants of fuel consumption and greenhouse gas (GHG) emissions). In fact, household motorized vehicle use accounts for about 58% of the GHG emissions from the transportation sector, a sector which itself is the single largest contributor to anthropogenic GHG emissions in the U.S. (EPA, 2021a, page ES 27). This contribution from household vehicle use to GHG emissions, despite substantial improvements in fuel efficiency through technological advancements, may be explained primarily by the shift in the body type of vehicles owned by households from passenger cars to light duty trucks. In particular, while light duty trucks (including pick-up trucks, minivans, and sport utility vehicles) accounted for just about 20% of the new vehicle market 35 years ago and 50% of the fleet less than ten years back, they accounted for about 70% of the new vehicle market in 2020 (EPA, 2021b, page 21).

The importance of modeling and predicting household vehicle holdings by body type and use has not gone unnoticed by travel demand researchers and practitioners. Specifically, while much of the research in the past in the area was focused on the body type of the most recently purchased or most used vehicle in a household, and even that in broad binary classes such as car versus non-car or sports utility vehicle (SUV) versus non-SUV vehicles, more recent studies have considered the entire household vehicle holdings using a more disaggregate classification of vehicle type (see Ma and Ye, 2019 for a review of such studies). Further, because of the increasing interest in, and legislative initiatives to, proactively influence the regional fleet mix of vehicles through environmental policies aimed at reducing mobile-source pollutants and GHG emissions, models of household vehicle fleet composition are being embedded within larger activity-based travel and emissions forecasting systems (see, for example, Vyas et al., 2012, You et al., 2014, and Cambridge Systematics, 2021).

A modeling framework that has received substantial attention in household vehicle fleet mix and use modeling is the MDCEV model, which incorporates the notion that households own and use different vehicles for different functional purposes. As such, the MDCEV model framework offers an elegant, theoretically consistent, and econometrically integrated approach to jointly model household vehicle ownership, vehicle type, and vehicle usage decisions. But the traditional MDCEV approach (for example, Garikapati et al., 2014 and Cambridge Systematics, 2021) needs an overall budget of miles traveled by all modes (including by non-motorized modes and other non-private vehicle modes) to be able to have the travel by motorized private vehicle modes to be sensitive to such characteristics as individual and household demographics, built environment, and fuel costs. This is achieved by assuming the presence of an outside good labeled, for example, as the "non-private mode". But since the mileage traveled by such "non-private modes" is not readily available from typical survey data, *ad hoc* assumptions are made for this outside good mileage. Our approach in this paper, however, does away with the need to make such *ad hoc* assumptions.

4.2. The Data and Sample Description

The sample for this demonstration is based on vehicle ownership and use data from the 2017 National Household Travel Survey in the state of Texas. The motorized private vehicles owned by each household are categorized into one of five vehicle types: (1) Passenger cars (coupes, sedans, hatchbacks, crossovers, and station wagons; cars for short), (2) Vans, (3) Sports Utility Vehicles (SUVs), (4) Pickup trucks (PUTs), and (5) Other (non-pickup trucks and recreational vehicles). The final estimation sample includes 1423 Texan households (including those with zero vehicle ownership) who owned no more than one vehicle within each of the five vehicle types. Of course, a household might own multiple vehicle types. A separate hold-out validation sample of 418 Texan households was also created (details of the sample formation are provided in Bhat et al., 2020, though the sample in the current study also includes households with zero vehicles that were not included in the earlier study). The MDC variable corresponds to ownership of each privately owned vehicle type and the amount of annual miles on each vehicle type (zero annual miles using private automobiles is allowed).

Table 1 provides information on the distribution of vehicle types in the vehicle-use dataset (the table summarizes the statistics across the estimation and validation samples, for a total of 1841 households). Not surprisingly, the percentage of zero-car households is quite low at 3.4% (see the last row under the column "zero vehicle HH"). Also, as expected, the most frequently owned vehicle types (if one or more vehicles are owned) correspond to passenger cars, SUVs, and PUTs (see the penultimate column of Table 1, which shows a total of 1138 households (61.8%) owning cars, 825 (44.8%) households owning SUVs, and 615 (33.4%) households owning PUTs). At the other end, vans and other types of vehicles (non-pickup trucks and recreational vehicles) are the least likely to be present in household vehicle fleets, with only 163 (8.9%) households owning vans and only 95 households (5.2%) owning non-pickup/recreational vehicles. The percentage of pickup trucks and vans in the mix increases within households with more than one vehicle. Across all households, the vehicle fleet in the sample includes 57% of light duty trucks (vans, SUVs, and PUTs), which is about the range of light duty trucks in the US household vehicle fleet mix about five years back. In terms of vehicle-use, the last column of Table 1 indicates that SUVs tend to be the most widely used if held by a household, followed by pickup trucks and passenger cars.

4.3. Model Specification

Several types of variables were considered in the first stage total annual mileage model as well as the second stage fractional MDCEV model. These included household sociodemographics (household size, presence and number of children, number of workers, household income, family structure, and dwelling type), and residential density and employment density variables.

The focus here is on demonstrating the application of the proposed model rather than necessarily on substantive interpretations and policy implications. But, we did undertake a rigorous specification analysis with the data available to arrive at the best possible specification (including considering alternative functional forms for continuous independent variables such as income, including a linear form, piecewise linear forms in the form of spline functions, and dummy variable specifications for different groupings). Further, to accommodate heterogeneity across individuals in the effect of observed variables not only in the baseline preference function (the ψ_k function as in Equation (4)), but also in the satiation parameters (the γ_k parameters), we parameterized the satiation parameters as $\exp(\delta_k' \omega_k)$, where ω_k is a vector of decision maker-related characteristics and δ_k is a vector to be estimated (note that $\gamma_k > 0 \; \forall \, k$). This allows the discrete choice decision of consuming an alternative (owning a particular vehicle type) to be less closely tied to the continuous choice of the amount of consumption (that is, vehicle mileage) of that alternative (see Bhat, 2008).

The performance of our two-stage linked Reverse Gumbel MDCEV (or simply the "linked" model for presentation convenience) is examined against a two-stage unlinked MDCEV model (or simply the "unlinked" model for presentation convenience). The latter unlinked model uses an independent Tobit model for modeling the total budget and an unlinked fractional MDCEV model given the budget. In this unlinked model, the fractional MDCEV component and the corresponding forecasting expressions remain the same as earlier, but Equation (26) now takes the following unlinked form:

$$\tilde{y}^* = \tilde{\boldsymbol{\theta}}'\mathbf{s} - \tilde{\lambda}\tilde{\zeta} \ \ , \ y \ = \begin{cases} 0 & \text{if } \ \tilde{y}^* \le 0 \\ \tilde{y}^* & \text{if } \ \tilde{y}^* > 0 \end{cases} \tag{41}$$

with $\tilde{\zeta}$ now being a reverse standard Gumbel variate, and $\tilde{\lambda}$ being a scale parameter. Of course, equivalently, the equation above may be written in terms of a traditional standard Gumbel variate υ as:

$$\tilde{y}^* = \tilde{\boldsymbol{\theta}}'\mathbf{s} + \tilde{\lambda}\upsilon \ \ , \ y \ = \begin{cases} 0 & \text{if } \ \tilde{y}^* \le 0 \\ \tilde{y}^* & \text{if } \ \tilde{y}^* > 0 \end{cases} \tag{42}$$

The likelihood component for an observation with no budget allocation to the product category in this unlinked case is:

$$L(\tilde{\boldsymbol{\mu}})\text{=Prob}(\tilde{y}^* < 0) = F_{\tilde{y}^*}\left(0\right) = \text{Prob}\left(\upsilon < -\frac{\tilde{\boldsymbol{\theta}}'\mathbf{s}}{\tilde{\lambda}}\right) = F_{\upsilon}\left(-\frac{\tilde{\boldsymbol{\theta}}'\mathbf{s}}{\tilde{\lambda}}\right), \text{ where } F_{\upsilon}\left(t\right) = e^{-e^{-t}} \ , \tag{43}$$

The corresponding likelihood component for an observation with non-zero budget allocation to the product category is:

$$L(\tilde{\boldsymbol{\mu}})\text{=}f_{\tilde{y}^*}(h) \times P\left(\tilde{f}_2^{op},...,\tilde{f}_M^{op},0,0,...,0\right), \quad \text{with } f_{\tilde{y}^*}(y \ ) = \tilde{\lambda}^{-1} e^{-e^{-\left[(y \ - \tilde{\boldsymbol{\theta}}'\mathbf{s})/\tilde{\lambda}\right]}} e^{-\left[(y \ - \tilde{\boldsymbol{\theta}}'\mathbf{s})/\tilde{\lambda}\right]}. \tag{44}$$

The moments of the resulting censored traditional Gumbel distribution for y in this unlinked model may be readily computed (interestingly, even the moments of censored versions of the traditional Gumbel distribution appear to have only been recently formally derived and presented; see Baran et al., 2021 and Neamah and Qasim, 2021). With the results from these papers, the equivalent expressions to Equation (31) in this unlinked case (as used in model evaluation later) are:

$$\begin{aligned} E(y \ ) &= \left[E(\tilde{\boldsymbol{\theta}}'\mathbf{s} + \tilde{\lambda}\upsilon) \,|\, \left(\tilde{\boldsymbol{\theta}}'\mathbf{s} + \tilde{\lambda}\upsilon > 0\right)\right] \times P(\tilde{\boldsymbol{\theta}}'\mathbf{s} + \tilde{\lambda}\upsilon > 0) \\ &= \left[\tilde{\boldsymbol{\theta}}'\mathbf{s} - \left\{\left[1 - F_{\upsilon}\left(-\frac{\tilde{\boldsymbol{\theta}}'\mathbf{s}}{\tilde{\lambda}}\right)\right]^{-1} \tilde{\lambda}\delta\right\}\right] \times \left(1 - F_{\upsilon}\left(-\frac{\tilde{\boldsymbol{\theta}}'\mathbf{s}}{\tilde{\lambda}}\right)\right) = \tilde{\boldsymbol{\theta}}'\mathbf{s}\left(1 - F_{\upsilon}\left(-\frac{\tilde{\boldsymbol{\theta}}'\mathbf{s}}{\tilde{\lambda}}\right)\right) - \tilde{\lambda}\delta, \text{ and} \\ Var(y \ ) &= \left[Var(\tilde{\boldsymbol{\theta}}'\mathbf{s} + \tilde{\lambda}\upsilon) \,|\, \left(\tilde{\boldsymbol{\theta}}'\mathbf{s} + \tilde{\lambda}\upsilon > 0\right)\right] \times P(\tilde{\boldsymbol{\theta}}'\mathbf{s} + \tilde{\lambda}\upsilon > 0). \\ &= \left(\left(\tilde{\boldsymbol{\theta}}'\mathbf{s}\right)^2 - \left\{\left[1 - F_{\upsilon}\left(-\frac{\tilde{\boldsymbol{\theta}}'\mathbf{s}}{\tilde{\lambda}}\right)\right]^{-1} \left[\tilde{\lambda}^2\delta^2 - 2\tilde{\lambda}\left(\tilde{\boldsymbol{\theta}}'\mathbf{s}\right)\right]\right\}\right) - \left[\tilde{\boldsymbol{\theta}}'\mathbf{s} - \left\{\left[1 - F_{\upsilon}\left(-\frac{\tilde{\boldsymbol{\theta}}'\mathbf{s}}{\tilde{\lambda}}\right)\right]^{-1} \tilde{\lambda}\delta\right\}\right]^2, \end{aligned} \tag{45}$$

where $F_{\upsilon}\left(t\right)=e^{-e^{-t}}$, $\delta = \int\limits_{t=0}^{t=\exp(\tilde{\boldsymbol{\theta}}'\mathbf{s}/\tilde{\lambda})} \exp(-t)(\ln t)dt$, and $\tilde{\delta} = \int\limits_{t=0}^{t=\exp(\tilde{\boldsymbol{\theta}}'\mathbf{s}/\tilde{\lambda})} \exp(-t)(\ln^2 t)dt$.

We undertake a comparative data fit investigation of the linked and unlinked models both in the estimation sample as well as the hold-out sample, and using both likelihood-based measures as well as more intuitive non-likelihood based measures. Because the direction of effects of variables from the two models were similar, we will not present the results of the unlinked model in this paper (this is available from the author). However, some of the substantive differences between the linked and unlinked models are discussed, even as our focus will be more on the data fit measures between our proposed model and the unlinked model.

4.4. Model Results

For completeness, we now discuss the substantive results from our proposed linked model. Table 2 presents the results for the fractional MDCEV model, and Table 3 presents the results for the Tobit budget model.

4.4.1. MDCEV Fractional Split Model Results

The parameter estimates in Table 2 relate to the impact of variables on the logarithm of the baseline preference, except for the results specific to satiation toward the bottom of the table. The five vehicle type alternatives are (1) Passenger car, (2) Van, (3) SUV, (4) PUT, and (5) Other. In the table, we retain some variables whose coefficients do not rise to a statistical significance level of 0.05, but that still provide intuitive and useful insights.

Household sociodemographic effects: Household demographics have a significant effect on vehicle type choice, particularly the effect of annual household income. Table 2 indicates that low income households (less than $35,000 annual income) are more likely to own vans and less likely to own “other” (non-pickup and recreational) vehicles, while increasingly higher income households are more likely to own SUVs. This latter result is not surprising, because most of the luxury vehicles reside within the SUV class, and SUVs are known to be gas-guzzlers that require quite a bit of fuel cost outlays. Interestingly, the preference for PUTs is highest in the middle income bracket ($75-$125K annual income range), though households with higher than $35K annual income more generally have a higher preference for PUTs than those with household income less than $35K. The preference for PUTs in the middle-income bracket may be reflective of the use of such vehicles for work-related activities associated with farming and transportation.

In terms of household lifecycle effects, there is a clear preference for vans among households with children (less than 15 years of age), presumably because vans are more spacious, safe, and comfortable for travel with small children. Besides, parents may prefer a van also because it opens up the possibility of carpooling arrangements of children with other parents, making the transportation of children efficient across multiple families and engendering a mutually beneficial arrangement for all parents involved. In addition to the effect of children on the preference for vans, the results also indicate that households with more adult individuals prefer vans to other vehicle types, and least prefer cars. Interestingly, after controlling for number of adults and children, there was no additional effect of number of workers in the household on vehicle type choice. This is, in part, because of a rather high correlation between number of adults and number of workers within the sample of Texan households used here, as well as because vehicle types other than cars have become so mainstream in the vehicle fleets of households, and serve distinct functionalities in the everyday lives of individuals, that employment status of individuals does not play much of a role in vehicle type choice. Finally, within the category of household demographics, the race of the individuals in a household is also found to impact vehicle-type holding and usage, even after controlling for income effects. Specifically, all-white households are clearly much more likely to own pickup trucks and other vehicle types (other truck types/recreational vehicles). As also pointed out by Bhat et al. (2020), pick-up truck ownership statistics do reveal that about 75% of the purchases of the top five pickup trucks in the U.S. are by white households.

Household and work location attribute effects: Relative to households in areas with a low population density (4000 persons per square mile or lower in the census block group of the household's residence), households in locations with high population density (more than 4000 persons per square mile in the Census block group of the household's location) have a higher preference for passenger cars. This result is to be expected, reflecting the relative ease of maneuverability with small-sized vehicles in highly dense travel areas as well as the higher fuel efficiency afforded by cars in stop-and-go traffic. The result regarding the higher inclination of households residing in less dense employment locations (less than 750 employees per square mile) to own pickup trucks supports the earlier suggestion that such households may be more likely to be self-employed in farming and other related pursuits, and trucks make it particularly convenient to haul large-sized items and operate in relatively rugged terrain.

Baseline preference constants: The presence of count variables (such as number of adults and children) in the specification renders the interpretation of the baseline preference constants difficult. However, as expected, the vans and other (non-pickup trucks/recreational) vehicle types have the highest negatively valued constants, conforming to their relatively low representation in household vehicle fleets compared to cars, SUVs, and PUTs.

Satiation effects through γ_k parameters: The satiation parameters (the γ_k parameters) are parameterized as $\exp(\boldsymbol{\delta}_k'\boldsymbol{\omega}_k)$, and the results in the lower panel of Table 2 represent the elements of the coefficient vector $\boldsymbol{\delta}$. A positive coefficient has the effect of increasing the γ_k parameter and, thereby, reducing satiation effects, while a negative coefficient has the effect of decreasing the γ_k parameter and increasing satiation effects. The results reveal that, conditional on ownership, households with an annual income over 35K tend to put less mileage (higher satiation) on PUTs and other recreational vehicles compared to other vehicles. Given that most households in this range are likely to have a combination of passenger cars and PUTs, it stands to reason that these households will put more mileage on the smaller vehicle from a fuel cost standpoint. In addition, the results also suggest that, conditional on ownership, the highest income households will put less mileage on SUVs (relative to other vehicle types) than their lower-income peers. While this may seem counter-intuitive, it is simply the model's way of reflecting the reality that, should households in the low and middle-income ranges happen to decide to own SUVs (which must mean that they are likely to particularly value the functionality of SUVs, given that households in these income ranges are not likely to own SUVs in the first place), the intensity of SUV use in such SUV-owning low-to-middle income households may not be much different from the intensity of SUV use in SUV-owning high income households. Thus, given the high baseline preference for SUVs among households in the highest income bracket, and the fact that the baseline preference not only dictates the discrete consumption choice, but also serves as the basis from which satiation effects start operating, the model increases the satiation parameter for high income households to ensure that SUV use among SUV-owning households do not vary much based on income earnings. Interestingly, though, the situation is reversed for vans, where higher income households, if they choose to own vans, appear to use such vehicles quite considerably. This may reflect the association between choosing vans and wanting to pursue leisure trips among high income households. Noteworthy in the results for the satiation parameters is that no other household or individual or location attribute affects intensity of use. This is evidence of the close association between vehicle type purchase and use decisions. That is, the intended use intensity and functionality of vehicle types is carefully considered by households even as a purchase is being made, rather than households determining intensity of use after purchasing vehicles. The constants related to the satiation parameters (the last row of Table 2) again have no clear interpretation; they work alongside the baseline parameters to determine intensity of usage after accounting for observed variable effects.

4.4.2. Total Mileage Tobit Model Results

The coefficients in Table 3 provide the Tobit model results. The coefficients in the table correspond to the effects of variables (elements of the $\boldsymbol{\theta}$ vector) and the coefficient λ (linking parameter) on the underlying latent propensity corresponding to the total intensity of travel (that is, total motorized mileage), as in Equation (26). The results here are intuitive, indicating the lower travel mileage propensity among (a) low income households (annual income of less than \$35K), (b) households with more adults and more workers, (c) white households, and (d) households located away from high population and employment density locations.

Most importantly, in the context of the current paper, the linkage parameter turns out to be 0.713. The standard error is 0.0155, with a corresponding highly statistically significant t-statistic of 21.83 relative to the value of 0 (note that, unlike a log-sum in a nested logit model, there is no interpretation for the linking parameter to have the value of 1, and there is no requirement that the parameter value must fall in the range of 0-1; in our application, it so happens to be in this range). The linkage result is clear indication of the endogeneity of the total mileage to the composition of the household vehicle fleet, and reflects the strong effects of variables impacting vehicle type decisions on the total mileage. Of course, the linkage parameter, as already discussed, also reflect scale effects in the Tobit regression, and generates heteroscedasticity in the underlying y^* variable, as should be clear from the expression for the variance of the error term η (see Equation 30) embedded in y^*.[14] In this context, in the unlinked model, the error term for the underlying latent variable for the Tobit regression (that is, the error term in $\tilde{y}^*$, which is $\tilde{\lambda}\tilde{\zeta}$ in Equation (41)) is homoscedastic. As is well known, ignoring heteroscedasticity when present in a Tobit model will, in general, lead to inconsistent estimation, which is another problem with the unlinked model.

But accommodating linkage is not simply an esoteric econometric issue. It can have important policy implications from a structural standpoint. For example, based on our results, neighborhood densification (through employment densification) would reduce motorized travel directly (based on the negative sign in the Tobit regression on employment density), but also have an additional indirect negative linkage effect though the decrease in PUT baseline preference in high employment density areas (based on the negative sign on the PUT baseline preference in Table 2). That is, the results show that densification will have a more negative effect on total motorized mileage (through the cumulative of the two combined effects just discussed) than what would be estimated if the linkage were not considered. In addition, in our empirical analysis, the direct effect of employment density is –0.323, while the unlinked Tobit model indicated a complete lack of an employment density effect (the coefficient was not statistically different from zero at even the 0.27 level of significance). The net result is that the linked model estimates a clear reduction in total motorized mileage overall due to densification, while the unlinked model indicates no such reduction. Further, the reduction in total mileage from the linked model then leads to a large reduction in PUT mileage in the linked model.

Finally, we should also note that the introduction of the linkage will, in general, provide more stability in estimation by adding a non-linear (in components) linking term with good variation in its values. For example, while in the linked model, the coefficient on the variable "income less than $35K" turned out to be –0.587 and highly statistically significant, the corresponding coefficient in the unlinked model turned out to have a large standard error, with associated convergence problems in estimation.

4.5. Data Fit Measures

Data fit measures are presented in two forms – likelihood-based data fit measures and non-likelihood based data fit measures.

4.5.1. Likelihood-Based Data Fit Measures

In the estimation sample, we estimate a base model with four constants in the baseline preference (number of alternatives minus 1), five constants for the satiation effects (one per alternative), a

[14] The heteroscedasticity depends upon the a_k values in a complicated, but utility theoretic, manner.

constant in the Tobit model, and the scale (in the linked model, the scale and the linking parameters get confounded and cannot be disentangled, as discussed earlier). The base model will not have the same log-likelihood at convergence for the linked and the unlinked models because the distribution of the kernel stochastic term is different between the two models (reverse Gumbel in the unlinked model and minLogistic in the linked model). But we do compute a $\bar{\rho}^2$ value for each of the fully specified linked and unlinked models relative to the base model for the unlinked specification.

$$\bar{\rho}^2 = 1 - \frac{L(\hat{\boldsymbol{\theta}}) - M}{L(c)}, \tag{46}$$

where $L(\hat{\boldsymbol{\theta}})$ and $L(c)$ are the log-likelihood functions at convergence and the base unlinked model, respectively, and M is the number of parameters (excluding the constants) estimated in the model. If the difference in the indices is $(\bar{\rho}_2^{\,2} - \bar{\rho}_1^{\,2}) = \tau$, then the probability that this difference could have occurred by chance is no larger than $\Phi\{-[-2\tau L(c) + (M_2 - M_1)]^{0.5}\}$, with a small value for the probability of chance occurrence suggesting that the difference is statistically significant and the model with the higher value for the adjusted likelihood ratio index is preferred.

All of the above metrics correspond to the estimation sample. We next undertake a similar analysis for the hold-out sample, maintaining the estimated coefficients. These predictive likelihood-based measures for the unlinked and linked models may be informally compared in terms of $\bar{\rho}^2$ fit.

The likelihood based data fit measures for the estimation sample are provided in Table 4. Both the linked and unlinked model log-likelihood values are clearly superior to the base "constants only" model, as can be observed from the nested likelihood ratio test (fifth row) for the estimation sample). These results demonstrate the value of our variable specification. Also, from the non-nested likelihood ratio statistics value provided in the final row, it can be inferred that the probability of the adjusted likelihood ratio index difference between the linked and the unlinked model occurring by chance is literally zero. The superior fit of the linked model carries over to the hold-out sample, with the $\bar{\rho}^2$ measure being 0.055 for the linked model and 0.049 for the unlinked model. Overall, the likelihood measures clearly favor the linked model over the unlinked model in the current empirical context.

4.5.2. Non-Likelihood Fit Measures

To further supplement the disaggregate likelihood-based performance at the multivariate and disaggregate levels, we use the hold-out sample to evaluate the performance of the models intuitively and informally at a disaggregate and aggregate level. At the disaggregate level, we estimate the probability of the observed multivariate discrete-continuous outcome for each individual, and compute an average probability of correct prediction for the consumption outcome. At the aggregate level, to keep the presentation manageable, we focus on only those households that own two vehicles. This also will be useful to assess the performance when households hold more than one vehicle, which is the reason to use an MDC-based model in the first place. We then design an informal heuristic diagnostic check of model fit by computing the aggregate percentage of households predicted to hold each of the 10 vehicle type (paired) combination outcomes. These predicted percentages falling into each combination category are compared with the actual percentage of households in each combination (using both a mean absolute percentage error (MAPE) statistic and a weighted mean absolute percentage error (MAPE) statistic, which is the MAPE for each combination weighted by the actual percentage shares of households in each combination).

For the continuous consumption predictions, we predict the continuous mileages for each household and each vehicle among households predicted to have two vehicles, using $\hat{y}_k = \tilde{f}_k^{op} E(y)$ from Step 9 of the forecasting algorithm. In using this procedure, we use 1000 error vector replications per individual observation. We then compute the aggregate predicted continuous mileage values for each vehicle type across these households, and compare the predicted mileages against the actual mileages by vehicle type for two-vehicle owning households.

In terms of the results, the average probability of correct prediction (APCP) at the multiple discrete-continuous level (across all combinations, including zero vehicles) is 0.0635 for the unlinked model and 0.0660 for the linked model. This difference across all combinations is marginal, and to be expected because the fractional MDCEV model is not affected much by linkage or no-linkage. Where we can expect more difference is in the prediction of zero vehicle count (because the Tobit model is the one that determines zero vehicle count) and the continuous consumption values (also determined by the Tobit model, as this model provides the overall mileage across all motorized vehicles). In this regard, the APCP for the discrete outcome of zero vehicle households is 0.2326 for the unlinked model, but increases to 0.2485 for the linked model (in the current empirical setting, this improvement of the linked model for zero-vehicle households gets tempered when the average probability of correct predictions is considered across all vehicle type combinations, because of the low percentage of households with zero vehicles in Texas). Within households with non-zero vehicles, the APCP of the actual vehicle combination owned and the observed vehicle mileage on each owned vehicle is 0.0572 for the unlinked model and 0.0592 for the linked model.

Moving on to the predictions of vehicle type holdings in two-vehicle households, the left panel of Table 5 provides the observed and predicted percentage of households in each possible two-vehicle combination. In the three combinations that make up over 84% of households (corresponding to the Car-SUV, the Car-Pickup, and the SUV-Pickup combinations), the linked model clearly outperforms the unlinked model. This is also reflected in the weighted mean absolute percentage error (MAPE), which is 23.2% for the unlinked model, but only 16.0% for the linked model (see the last row of the table) across all the ten combinations for two-vehicle households).[15] The right side panel of Table 5 provides the mean observed continuous-level mileages for each vehicle type within each combination (the mean being computed across all households falling in each combination), and the corresponding mean continuous-level predictions (the mean being computed across households predicted to fall in each combination). The table does not show the results for the combinations including the "other" vehicle type because of the extremely small fractions of households in these combinations and the wide variations in mileage across the small number of households with "other" vehicle usage). Several noteworthy observations may be made based on these results. First, the predicted mean mileages are remarkably close to the actual mean mileages for cars, the most often owned vehicle type by two-vehicle households, with the APE for car mileage being less than 15% for all combinations that include a car. Second, the linked model does slightly better than the unlinked model on both the unweighted and weighted mean MAPE values. Third, unlike the case of discrete consumptions where the weighted performance of the linked model is substantially superior to that of the unlinked model, there is a smaller difference in the accuracy of the continuous mileage predictions in the current empirical context.

Overall, however, the performance of the proposed model at the aggregate as well as disaggregate levels reinforces the notion that the choice of vehicle type combination, the intensity of use of each vehicle type, and the total mileage are all closely linked, and emphasizes the value of modeling these dimensions based on the linked structure proposed in this paper. Importantly, given the different stochastic distributional forms of the linked and unlinked models, data fit alone need not be the guiding factor in choosing the linked model (in fact, it is not inconceivable that the unlinked model would perform even better than the linked model from a pure data fit standpoint in some empirical situations). But, from a behavioral standpoint, the linked model accommodates the notion that changes in the attributes of the goods within the product group of interest not only have a substitution influence, but also an income effect through a change in the total consumption quantity in the product group. As we have shown in Section 2.2.1, this linkage between inside good attributes and the total consumption quantity in the product group is consistent with two-stage budgeting and utility maximization, while the linked model completely ignores such linkage and is not consistent with utility maximization. Thus, regardless of data fit considerations, in most circumstances, the analyst may prefer to use the linked model from a behavioral standpoint.

5. CONCLUSIONS

In this paper, we propose an approach that does not need the total budget to be observed or predetermined, allows for any finite or not-so-finite budget over the entire set of inside and outside goods, and preserves a strong endogenous utility-theoretic link between inside good consumptions and the budget allocated to the inside goods (that is, to the product group of interest). At the same time, the proposed approach also makes the forecasting process simple and

[15] We do not consider the MAPE for the "van-other" combination because there were no households in this combination in the hold-out sample.

easy. Our approach, inspired by similar efforts in the past in the context of the vertical choice-making approach, models an endogenous budget for the inside goods as well as consumption of the inside goods separately, but within a unified utility-theoretic framework. As importantly, different from earlier applications in the vertical choice making approach, we also allow unobserved heterogeneity in the intensity of linking between the inside good preferences and the budget allocation for the product group of interest. We show that our proposed model, including a fractional MDCEV model at the lower level linked up to a Tobit model for the budget allocation to the inside goods, is strictly consistent with a two-stage budgeting utility theoretic structure. Then, by using a reverse Gumbel distributional assumption for the stochastic terms in the baseline preferences of each of the inside alternatives in the fractional MDCEV model, and a reverse Gumbel distribution for the random error term in the Tobit model, we derive an incredibly simple closed-form model for the resulting multiple discrete-continuous extreme value model that, to our knowledge, is a first of its kind in the econometric literature. In doing so, we formally introduce a new distribution, which we label as the minLogistic distribution, to the statistical literature, and derive the properties of the distribution that is then used in the forecasting stage of the proposed model. An application of the proposed model to investigate the household vehicle fleet composition and usage demonstrates its potential relative to an unlinked and exogenously developed budget for the inside goods.

Of course, there are many directions along which the proposed model may be extended, most of which also are certain to dismantle the closed-form nature of the proposed model. But, with strategic accommodation, some of these extensions should be readily estimable because of advances in simulation and analytic approximation techniques. First, as discussed in Section 3, random parameters may be added to one or both of the fractional MDCEV model or the Tobit model, which can allow for unobserved heterogeneity in sensitivity to exogenous variables, as well as correlation across the inside good preferences. Second, the Tobit model may be replaced by a system of two equations, one equation for zero versus positive total consumption for the inside goods and another for the total budget allocation to the inside goods, given positive consumption. The linking term from the fractional MDCEV can be introduced in each of these equations, while still retaining a closed-form model. Of course, one can further introduce a correlation across the two equations (that may replace the Tobit model) through an error-component mixing approach, or by employing well-known bivariate parametric distributions. Third, from an empirical standpoint, while the traditional single-stage budgeting MDCEV model is not applicable for cases with an unknown total budget over the inside and outside goods (our proposed model is), it would be interesting to undertake an empirical comparison of the traditional MDCEV model with the proposed two-stage budgeting MDCEV model for cases when the budget is observed. Both models are applicable in this situation. While the proposed model does have the benefit of disentangling substitution and income effects (see Section 1.1), a comprehensive empirical comparison of the two models based on data fit as well as policy implications may bring out interesting results. Fourth, approaches that relax the need for strict exogeneity of the linking function in the first stage budget model and that allow the first stage total budget itself to impact the second stage fractional splits would be helpful, though would almost definitely also destroy the utility-theoretic and/or closed-form nature of the model.

The proposed two-stage budgeting MDCEV-Tobit should prove to be beneficial in a number of multiple discrete-continuous choice contexts. The closed-form probability structure makes the estimation procedure no more difficult than for traditional MDCEV models. As such, we believe that the proposed model can open up a whole new world of MDC applications in general, particularly for those cases with an unobserved total budget over the inside and outside goods and/or general nested linkages of model systems with an MDC model at the lower level. In closing, we are excited by the prospect that the proposed model can add to the arsenal (intended and used here in only a pacifist way) at the disposal of choice modelers and econometricians when analyzing MDC situations.

ACKNOWLEDGMENTS

The author is grateful to Lisa Macias for her help in formatting this document. Vivek Verma helped verify, type up, and format the many integration formulas in the Appendix. The author would like to thank an anonymous reviewer for valuable comments.

REFERENCES

Anastasopoulos, P.Ch., Shankar, V.N., Haddock, J.E., Mannering, F.L., 2012. A multivariate tobit analysis of highway accident-injury-severity rates. *Accident Analysis & Prevention*, 45, 110-119.

Augustin, B., Pinjari, A.R., Eluru, N., and Pendyala, R.M., 2015. Estimation of annual mileage budgets for a multiple discrete-continuous choice model of household vehicle ownership and utilization. *Transportation Research Record: Journal of the Transportation Research Board*, 2493, 126-135.

Baran, S., Szokol, P., and Szabó, M., 2021. Truncated generalized extreme value distribution-based ensemble model output statistics model for calibration of wind speed ensemble forecasts. *Environmetrics*, 32(6), e2678.

Bhat, C.R., 2005. A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B*, 39(8), 679-707.

Bhat, C.R., 2008. The multiple discrete-continuous extreme value (MDCEV) model: Role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B*, 42(3), 274-303.

Bhat, C.R., 2018. A new flexible multiple discrete-continuous extreme value (MDCEV) choice model. *Transportation Research Part B*, 110, 261-279.

Bhat, C.R., and Pinjari, A., 2014. Multiple discrete-continuous choice models: A reflective analysis and a prospective view. In *Handbook of Choice Modelling* (Hess, S., and Daly, A., Eds) Chapter 19, 427-453, Edward Elgar Publishing Ltd.

Bhat, C.R., and Sen, S., 2006. Household vehicle type holdings and usage: An application of the multiple discrete-continuous extreme value (MDCEV) model. *Transportation Research Part B*, 40(1), 35-53.

Bhat, C.R., Castro, M., and Khan, M., 2013. A new estimation approach for the multiple discrete-continuous probit (MDCP) choice model. *Transportation Research Part B*, 55, 1–22.

Bhat, C.R., Paleti, R., and Castro, M., 2015a. A new utility-consistent econometric approach to multivariate count data modeling. *Journal of Applied Econometrics*, 30(5), 806-825.

Bhat, C.R., Castro, M., and Pinjari, A.R., 2015b. Allowing for complementarity and rich substitution patterns in multiple discrete-continuous models. *Transportation Research Part B*, 81(1), 59-77.

Bhat, C.R., Astroza, S., and Bhat, A.C., 2016. On allowing a general form for unobserved heterogeneity in the multiple discrete-continuous probit model: Formulation and application to tourism travel. *Transportation Research Part B*, 86, 223-249.

Bhat, C.R., Mondal, A., Asmussen, K.E., and Bhat, A.C., 2020. A multiple discrete extreme value choice model with grouped consumption data and unobserved budgets. *Transportation Research Part B*, 141, 196-222.

Bhat, C.R., Mondal, A., Pinjari, A.R., Saxena, S., and Pendyala, R.M., 2022. A multiple discrete continuous extreme value choice (MDCEV) model with a linear utility profile for the outside good recognizing positive consumption constraints. *Transportation Research Part B*, 156, 28-49.

Bockstael, N.E., Hanemann, W.M., and Kling, C.L., 1987. Estimating the value of water quality improvements in a recreational demand framework. *Water Resources Research*, 23(5), 951-960.

Cambridge Systematics, 2021. *New York best Practice Model 2012 Update Summary Report*. Prepared for New York Metropolitan Transportation Council (NYMTC).

Deaton, A., and Muellbauer, J., 1980. *Economics and Consumer Behavior*. Cambridge University Press, Cambridge.

Dziak, J.J., Coffman, D.L., Lanza, S.T., Li, R., and Jermiin, L.S., 2020. Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 21(2), 553-565

EPA, 2021a. Inventory of U.S. greenhouse gas emissions and sinks 1990-2019. Report EPA-430-R-21-005, United States Environmental Protection Agency.

EPA, 2021b. The 2021 EPA automotive trends report: Greenhouse gas emissions, fuel economy, and technology since 1975. Report EPA-420-R-21-023, United States Environmental Protection Agency.

Fang, H.A., 2008. A discrete-continuous model of households' vehicle choice and usage, with an application to the effects of residential density. *Transportation Research Part B*, 42(9), 736-758.

Garikapati, V.M., Sidharthan, R., Pendyala, R.M., and Bhat, C.R., 2014. Characterizing household vehicle fleet composition and count by type in integrated modeling framework. *Transportation Research Record: Journal of the Transportation Research Board*, 2429, 129-137.

Gorman, W.M., 1959. Separable utility and aggregation. *Econometrica*, 27(3), 469-481. https://doi.org/10.2307/1909472.

Gorman, W.M., 1961. On a class of preference fields. *Metroeconomica*, 13(2), 53-56. https://doi.org/10.1111/j.1467-999X.1961.tb00819.x

Hausman, J.A., Leonard, G.K., and McFadden, D., 1995. A utility-consistent, combined discrete choice and count data model: Assessing recreational use losses due to natural resource damage. *Journal of Public Economics*, 56(1), 1-30.

Hendel, I., 1999. Estimating multiple-discrete choice models: An application to computerization returns. *Review of Economic Studies*, 66, 423-446.

Hou, Q., Huo, X., and Leng, J., 2020. A correlated random parameters tobit model to analyze the safety effects and temporal instability of factors affecting crash rates. *Accident Analysis & Prevention*, 134, 105326.

Kim, J., Allenby, G.M., and Rossi, P.E., 2002. Modeling consumer demand for variety. *Marketing Science*, 21, 229-250.

Lee, L-F and Pitt, M.M., 1986. Microeconometric demand systems with binding nonnegativity constraints: The dual approach, *Econometrica*, 54(5), 1237-1242.

Ma, J., and Ye, X., 2019. Modeling household vehicle ownership in emerging economies. *Journal of the Indian Institute of Science*, 99(4), 647-671.

Mäler, K.G., 1974. Environmental economics: A theoretical inquiry. *The Johns Hopkins University Press for Resources for the Future*, Baltimore, MD.

Mannering, F.L., and Hamed, M., 1990. Occurrence, frequency and duration of commuters' work-to-home departure delay. *Transportation Research Part B*, 24(2), 99-109.

Mondal, A., and Bhat, C.R., 2021. A new closed form multiple discrete-continuous extreme value (MDCEV) choice model with multiple linear constraints. *Transportation Research Part B*, 147, 42-66.

Morey, E.R., Rowe, R.D., and Watson, M., 1993. A repeated nested-logit model of Atlantic salmon fishing. *American Journal of Agricultural Economics*, 75(3), 578-592.

Neath, A.A. and Cavanaugh, J.E., 2012. The Bayesian information criterion: Background, derivation, and applications. *WIREs Computational Statistics*, 4(2), 199-203.

Neamah, M.W., and Qasim, B.A., 2021. A new left truncated Gumbel distribution: Properties and estimation. *Journal of Physics: Conference Series*, 1897, 012015.

Paleti, R., Bhat, C.R., Pendyala, R.M., Goulias, K.G., Adler, T.J., and Bahreinian A., 2014. Assessing the impact of transportation policies on fuel consumption and greenhouse gas emissions using a household vehicle fleet simulator. *Transportation Research Record: Journal of the Transportation Research Board*, 2430, 182-190.

Pinjari, A.R., Augustin, B., Imani, V.S., Eluru, N., and Pendyala, R.M., 2016. Stochastic frontier estimation of budgets for Kuhn–Tucker demand systems: Application to activity time-use analysis. *Transportation Research Part A*, 88, 117-133.

Rouwendal, J., and Boter, J., 2009. Assessing the value of museums with a combined discrete choice/count data model. *Applied Economics*, 41(11), 1417-1436.

Saumard, A., and Wellner, J.A., 2014. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8, 45-114.

Saxena, S., Pinjari, A.R., and Bhat, C.R., 2022a. Multiple discrete-continuous choice models with additively separable utility functions and linear utility on outside good: Model properties and characterization of demand functions. *Transportation Research Part B*, 155, 526-557.

Saxena, S., Pinjari, A.R., Bhat, C.R., and Mondal, A., 2022b. A flexible multiple discrete-continuous probit (MDCP) model: Application to analysis of expenditure patterns of domestic tourists in India. Technical paper, Department of Civil Engineering, Indian Institute of Science (IISc).

Srinivasan, S., and Bhat, C.R., 2006. A multiple discrete-continuous model for independent- and joint- discretionary-activity participation decisions. *Transportation*, 33(5), 497-515.

Strotz, R.H., 1957. The empirical implications of a utility tree. *Econometrica*, 25(2), 269-280. https://doi.org/10.2307/1910254.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24-36. https://doi.org/10.2307/1907382.

Truong, T.P., and Hensher, D.A., 2014. Linking discrete choice to continuous demand in a spatial computable general equilibrium model. *Journal of Choice Modelling*, 12, 21-46.

Vasquez-Lavin, F., and Hanemann, M., 2008. Functional forms in discrete/continuous choice models with general corner solution. Department of Agricultural & Resource Economics, University of California Berkeley. CUDARE Working Paper 1078.

von Haefen, R.H., 2010. Incomplete demand systems, corner solutions, and welfare measurement. *Agricultural and Resource Economics Review*, 39(1), 22-36.

von Haefen, R.H., and Phaneuf, D.J., 2003. Estimating preferences for outdoor recreation: A comparison of continuous and count data demand system frameworks. *Journal of Environmental Economics & Management*, 45(3), 612-630.

Vyas, G., Paleti, R., Bhat, C.R., Goulias, K.G., Pendyala, R.M., Hu, H.-H., Adler, T.J., and Bahreinian, A., 2012. Joint vehicle holdings by type and vintage, and primary driver assignment model with application for California. *Transportation Research Record: Journal of the Transportation Research Board*, 2302, 74-83.

Wagner, J., Cook, J., and Kimuyu, P., 2019. Household demand for water in rural Kenya. *Environmental and Resource Economics*, 74(4), 1563-1584.

Wales, T.J., and Woodland, A.D., 1983. Estimation of consumer demand systems with binding non-negativity constraints. *Journal of Econometrics*, 21(3), 263-285.

You, D., Garikapati, V.M., Pendyala, R.M., Bhat, C.R., Dubey, S., Jeon, K., and Livshits, V., 2014. Development of vehicle fleet composition model system for implementation in activity-based travel model. *Transportation Research Record: Journal of the Transportation Research Board*, 2430(1), 145-154.

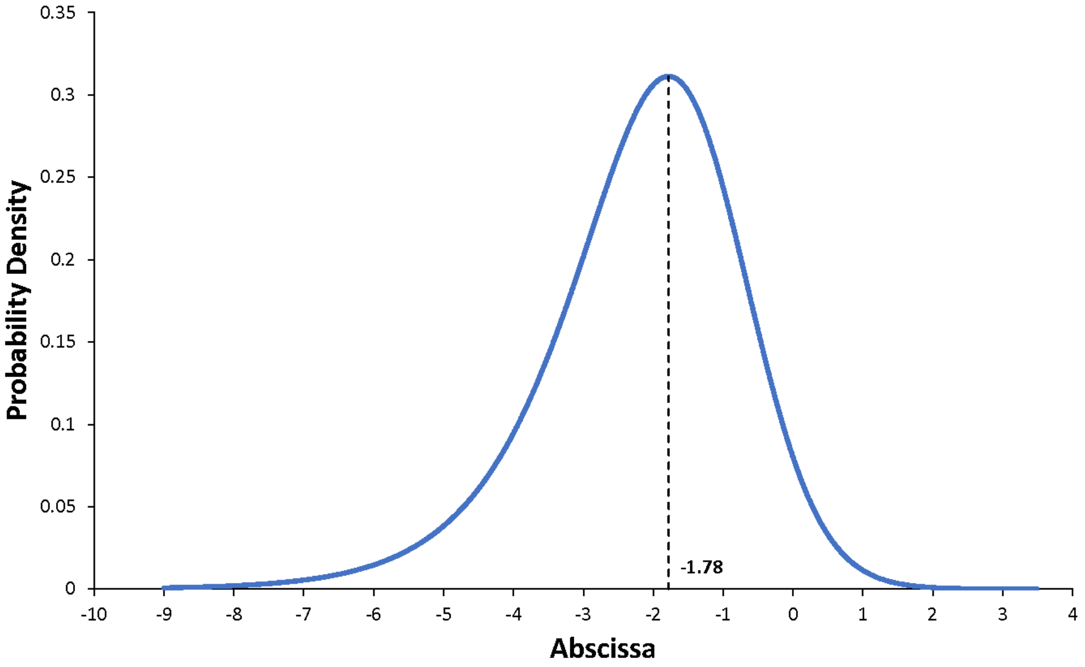

Figure 1. MinLogistic Distribution Density Function

Table 1. Data Description for the Vehicle-Use Case Study (sample size = 1841)

Vehicle-type	***Vehicle-type distribution***						
	Household (HH) vehicle ownership levels						**Average annual mileage**
	Zero vehicle HH	1-vehicle HH	2-vehicles HH	3-vehicles HH	4 or more vehicles HH	Total vehicles of each type	
Passenger car	--	490 (55.9%)	528 (34.7%)	106 (27.8%)	14 (24.1%)	1,138	8620
Van	--	37 (4.2%)	97 (6.4%)	22 (5.8%)	7 (12.1%)	163	7520
SUV	--	254 (29.0%)	463 (30.5%)	96 (25.2%)	12 (20.7%)	825	9895
Pickup truck	--	92 (10.5%)	404 (26.6%)	106 (27.8%)	13 (22.4%)	615	8805
Other	--	4 (0.4%)	28 (1.8%)	51 (13.4%)	12 (20.7%)	95	3740
Total	63 (3.4% of all households)	877 (100%)	1520 (100%)	381 (100%)	58 (100%)	--	--

Table 2. RG-MDCEV Fractional Split Model Results

Variables	Coefficient estimates (t-stats)				
	Passenger car	Van	SUV	Pickup truck	Other
Household sociodemographic					
Household Income					
Income less than $35,000 annually	--	0.331 (2.37)	--	--	-0.221 (-1.30)
Income between $35,000 - $75,000 annually	--	--	0.244 (2.14)	0.227 (1.71)	--
Income between $75,000 - $125,000 annually	--	--	0.478 (3.63)	0.590 (3.94)	--
Income greater than $125,000 annually	--	--	1.231 (7.26)	0.318 (1.95)	--
Number of children in the household	--	0.329 (6.58)	--	--	--
Number of adults in the household	-0.406 (-5.87)	0.317 (2.91)	--	--	--
Race is white (Base: Non-white)	--	--	--	0.432 (3.85)	0.590 (2.73)
Household location attributes					
Population density more than 4000 persons/sq. mile (Base: less than 4000 persons/sq. mile)	0.328 (3.75)	--	--	--	--
Employment density more than 500 workers/sq. mile (Base: less than 500 workers/sq. mile)	--	--	--	-0.323 (-3.38)	--
Baseline preference constants	--	-3.192 (-13.72)	-1.494 (-10.29)	-1.772 (-9.45)	-3.095 (-12.77)
Satiation effects					
Income between $35,000-$75,000 annually	--	--	--	-0.824 (-2.78)	-2.628 (-4.48)
Income between $75,000-$125,000 annually	--	0.724 (2.24)	--	-1.363 (-4.22)	-1.760 (-2.60)
Income greater than $125,000 annually	--	1.614 (4.67)	-2.142 (-7.32)	-0.837 (-2.08)	-3.781 (-6.28)
Satiation constant	-0.266 (-2.70)	1.994 (4.27)	0.942 (4.95)	0.575 (2.36)	0.988 (1.89)

Table 3. Tobit Model for Total Mileage by Motorized Modes

Variables	Coefficient estimates	t-statistics
Household sociodemographic		
Income less than $35,000 annually	-0.587	-6.06
Number of adults	0.180	3.00
Number of workers	0.383	9.79
Race is white	0.236	3.42
Household location attributes		
Population density more than 4000 persons/sq. mile (Base: less than 4000 persons/sq. mile	-0.088	-1.30
Employment density more than 500 workers/sq. mile (Base: less than 500 workers/sq. mile)	-0.264	-3.69
Linkage Parameter	0.713	21.83
Constant	0.182	1.09

Table 4. Likelihood Based Data Fit Measures

	Estimation Sample (N = 1423)	
	Unlinked Model	Linked Model
Log-likelihood at Convergence (predictive for hold-out sample)	-4,879.0	-4,868.1
Log-likelihood at Constants (predictive for hold-out sample)	-5,306.6	-5,336.3
Number of non-constant and non-scale parameters	41	41
Number of constant parameters (in MDCEV baseline preference and satiation, and Tobit model) plus scale parameter	11	11
Nested Likelihood Ratio Test w.r.t Constants Only Model (informal test in hold-out sample)	855	936
Adjusted Likelihood Ratio Index (predictive for hold-out sample) with respect to the unlinked base model	0.075	0.077
Non-Nested Likelihood Ratio Test between the Unlinked and Linked Models (informal test in hold-out sample)	Φ(-4.58) << 0.001	

Table 5. Aggregate Level Predictions for Two-Vehicle Households

	Discrete-level Prediction (Percentage)					Continuous-level Prediction (Annual Mileage Values are miles/1,000)											
		Unlinked Model		Linked Model		Unlinked Model							Linked Model				
						Vehicle 1			Vehicle 2			Within combination mileage MAPE	Vehicle 1		Vehicle 2		Within combination mileage MAPE
Combination	Actual	Predicted	APE	Predicted	APE	Actual	Predicted	APE	Actual	Predicted	APE		Predicted	APE	Predicted	APE	
Car-Van	6.59	4.87	26.1	4.66	29.2	9.68	8.38	13.4	13.29	8.53	35.8	23.0	7.75	19.9	9.33	25.6	23.3
Car-SUV	32.97	38.07	15.5	36.72	11.4	10.05	8.94	11.0	13.77	8.74	36.5	25.2	9.03	10.1	8.77	36.3	25.3
Car-Pickup	25.27	29.44	16.5	27.11	7.3	11.02	9.97	9.5	10.79	7.53	30.2	19.8	9.70	12.0	7.44	31.0	21.4
Car-Other	1.10	2.20	100.0	2.25	100.5	–	–	–	–	–	–	–	–	–	–	–	–
Van-SUV	1.65	2.41	46.1	2.48	50.3	15.13	8.89	41.2	12.33	10.16	17.6	30.6	11.90	21.3	8.81	28.5	24.6
Van-Pickup	3.30	2.64	20.0	2.49	24.5	9.88	10.27	3.9	4.30	8.45	96.5	32.0	11.27	14.0	8.03	86.7	36.1
Van-Other	0.00	0.40	–	0.40	–	–	–	–	–	–	–	–	–	–	–	–	–
SUV-Pickup	25.80	17.29	33.0	21.25	17.6	11.62	10.87	6.5	12.01	7.97	33.6	20.3	11.05	4.9	8.31	30.8	18.0
SUV-Other	1.10	1.39	29.0	1.37	24.5	–	–	–	–	–	–	–	–	–	–	–	–
Pick-Other	2.22	1.33	40.0	1.27	42.8	–	–	–	–	–	–	–	–	–	–	–	–
Mean APE	–	–	32.6%	–	31.8%	–	–	–	–	–	–	25.2%	–	–	–	–	24.8%
Weighted MAPE	–	–	23.2%	–	16.0%	–	–	–	–	–	–	22.6%	–	–	–	–	22.5%

APPENDIX A: MinLogistic Distribution Properties

A.1: Derivation of the Survival Distribution Function

From Equation (26),

$$S_\eta(t) = P\left[\left(\zeta - \sigma\left[\ln\sum_k a_k e^{\mu_k})\right]\right) > t\right]$$

$$= P\left[\left(\tilde{\zeta} - \left[\ln\sum_k a_k e^{\mu_k})\right]\right) > \tilde{t}\right], \text{with } \tilde{\zeta} = \frac{\zeta}{\sigma} \text{ and } \tilde{t} = \frac{t}{\sigma}$$

$$= P\left[\tilde{\zeta} > \left[\ln\sum_k a_k e^{\mu_k})\right] + \tilde{t}\right],$$

where $\tilde{\zeta}$ and μ_k are all standard (and independent) reverse-Gumbel terms. Using standard reverse-Gumbel distribution properties, we may write:

$$= \int\limits_{\mu_1=-\infty}^{+\infty} \int\limits_{\mu_2=-\infty}^{+\infty} \cdots \int\limits_{\mu_K=-\infty}^{+\infty} e^{-e^{\left[\tilde{t}+\ln\left(\sum_{k=1}^{K}\left[a_k e^{\mu_k}\right]\right)\right]}} e^{-e^{\mu_1}} e^{\mu_1} d\mu_1 e^{-e^{\mu_2}} e^{\mu_2} d\mu_2 \ldots e^{-e^{\mu_K}} e^{\mu_K} d\mu_K$$

$$= \int\limits_{\mu_1=-\infty}^{+\infty} e^{-e^{\tilde{t}}(a_1 e^{\mu_1})} e^{-e^{\mu_1}} e^{\mu_1} d\mu_1 \int\limits_{\mu_2=-\infty}^{+\infty} e^{-e^{\tilde{t}}(a_2 e^{\mu_2})} e^{-e^{\mu_2}} e^{\mu_2} d\mu_2 \ldots \int\limits_{\mu_K=-\infty}^{+\infty} e^{-e^{\tilde{t}}(a_K e^{\mu_K})} e^{-e^{\mu_K}} e^{\mu_K} d\mu_K$$

Next, consider the first integral. Straightforward integration and application of the limits provides the result that it is equal to $\frac{1}{(1+a_1 e^{\tilde{t}})}$. Other subsequent integrals may be similarly computed to give the result that:

$$S_\eta(t) = \frac{1}{\prod_{k=1}^{K}\left(1+a_1 e^{\tilde{t}}\right)} = \frac{1}{\prod_{k=1}^{K}\left(1+a_1 e^{t/\sigma}\right)}, \text{ as in Equation (27).}$$

Equation (28) results directly from the above survival distribution function.

A.2: Unimodality and Modal Value of the MinLogistic Distribution

To prove that the minLogistic distribution is unimodal, it suffices to show that the density function is log-concave. That is, that the logarithm of the density function is globally concave (see Saumard and Wellner, 2014).

Thus, we need to show that:

$$\frac{\partial^2 \ln f_\eta(t)}{\partial t^2} < 0$$

From Equation (28) in the text,

$$f_\eta(t) = \frac{1}{\sigma} \frac{e^{\frac{t}{\sigma}}}{\prod_{k=1}^{K}(1 + a_k e^{\frac{t}{\sigma}})} \sum_{k=1}^{K} \left(\frac{a_k}{1 + a_k e^{\frac{t}{\sigma}}} \right)$$

$$ln\, f_\eta(t) = -\, ln\, \sigma + \frac{t}{\sigma} - \sum_k ln\left(1 + a_k e^{\frac{t}{\sigma}}\right) + ln \sum_k \frac{a_k}{1 + a_k e^{\frac{t}{\sigma}}}$$

Defining $\xi_k = \dfrac{a_k e^{\frac{t}{\sigma}}}{1 + a_k e^{\frac{t}{\sigma}}}$; and after some straightforward differentiation, we get:

$$\frac{\partial\, ln\, f_\eta(t)}{\partial t} = \frac{1}{\sigma} - \frac{1}{\sigma}\left[\sum_{k=1}^{K} \xi_k + \frac{\sum_{k=1}^{K} \xi_k^2}{\sum_{k=1}^{K} \xi_k} \right]$$

Noting that $\dfrac{\partial \xi_k}{\partial t} = \xi_k^2 \dfrac{e^{-\frac{t}{\sigma}}}{\sigma a_k}$, and differentiating again, the end result is:

$$\frac{\partial^2\, ln\, f_\eta(t)}{\partial t^2} = -\frac{1}{\sigma^2} e^{-\frac{t}{\sigma}} \left[\frac{\left\{\sum_k \frac{\xi_k^2}{a_k}\right\} \{(\sum_k \xi_k)^2 - (\sum_k \xi_k^2)\} + 2(\sum_k \xi_k) \sum_k \left(\frac{\xi_k^3}{a_k}\right)}{(\sum_k \xi_k)^2} \right]$$

Because $\xi_k > 0, \left[\left(\sum_k \xi_k\right)^2 - \left(\sum_k \xi_k^2\right) \right] > 0,$

Which immediately implies that the second derivative above is always negative. Thus, $f_\eta(t)$ is unimodal with a unique mode. The mode does not have a closed form expression but can be obtained numerically by setting the first derivative to zero. That is,

$$\left[\sum_{k=1}^{K} \xi_k(\varpi) + \frac{\sum_{k=1}^{K} \xi_k^2(\varpi)}{\sum_{k=1}^{K} \xi_k\,(\varpi)} \right] = 1,$$

which is Equation (29) in the main text.

A.3: Expectation and Variance of Untruncated MinLogistic Distribution

$$f_\eta(t) = \frac{1}{\sigma} \frac{e^{\frac{t}{\sigma}}}{\prod_{k=1}^{K}(1 + a_k e^{\frac{t}{\sigma}})} \sum_{k=1}^{K} \left(\frac{a_k}{1 + a_k e^{\frac{t}{\sigma}}} \right)$$

1. $E(\eta) = \int_{-\infty}^{\infty} t f_\eta(t) dt = \int_{-\infty}^{\infty} t . \frac{1}{\sigma} \frac{e^{\frac{t}{\sigma}}}{\prod_{k=1}^{K}(1+a_k e^{\frac{t}{\sigma}})} \sum_{k=1}^{K} \left(\frac{a_k}{1+a_k e^{\frac{t}{\sigma}}} \right)$

$Substituting\; e^{\frac{t}{\sigma}} = u, t = \sigma \ln u, dt = \sigma . \frac{1}{u} du$, and rewriting,

$$\int_{-\infty}^{\infty} \frac{\sigma \ln u}{\sigma} \left\{ \frac{u\, a_1\, (1+a_2u)(1+a_3u)\ldots(1+a_Ku)}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2} + \frac{u\, a_2\, (1+a_1u)(1+a_3u)\ldots(1+a_Ku)}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2} + \cdots + \frac{u\, a_K\, (1+a_1u)(1+a_3u)\ldots(1+a_{K-1}u)}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2} \right\} \sigma.\frac{1}{u} du$$

$$= \sigma \int_0^{\infty} \ln u\,. \frac{a_1\, (1+a_2u)(1+a_3u)\ldots(1+a_Ku) + \ a_2\, (1+a_1u)(1+a_3u)\ldots(1+a_Ku) + \cdots + a_K\, (1+a_1u)(1+a_3u)\ldots(1+a_{K-1}u)}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2}$$

Integrating by parts:

$$Let\ I = \int fg' = fg - \int f'g$$

$$f = \ln u\,, f' = \frac{1}{u}$$

$$g' = \frac{a_1\, (1+a_2u)(1+a_3u)\ldots(1+a_Ku) + \ a_2\, (1+a_1u)(1+a_3u)\ldots(1+a_Ku) + \cdots + a_K\, (1+a_1u)(1+a_3u)\ldots(1+a_{K-1}u)}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2}$$

$$g = -\frac{1}{(1+a_1u)(1+a_2u)\ldots(1+a_Ku)}$$

$$So, I = -\frac{\ln u}{(1+a_1u)(1+a_2u)\ldots(1+a_Ku)} - \int -\frac{1}{u.(1+a_1u)(1+a_2u)\ldots(1+a_Ku)} du$$

$$Solving\ the\ Integral{:}\ J = \int \frac{1}{u.(1+a_1u)(1+a_2u)\ldots(1+a_Ku)} du$$

Performing partial fractions decomposition:

$$J = -\frac{a_1^K}{(a_1-a_2)(a_1-a_3)..(a_1-a_K)} \int \frac{1}{1+a_1u} du - \frac{a_2^K}{(a_2-a_1)(a_2-a_3)..(a_2-a_K)} \int \frac{1}{1+a_2u} du \ldots - \frac{a_K^K}{(a_K-a_1)(a_K-a_2)..(a_K-a_{K-1})} \int \frac{1}{1+a_Ku} du + \int \frac{1}{u} du$$

$$J = -\frac{a_1^{K-1}}{(a_1-a_2)(a_1-a_3)..(a_1-a_K)} \ln(1+a_1u) - \cdots - \frac{a_K^{K-1}}{(a_K-a_1)(a_K-a_2)..(a_K-a_{K-1})} \ln(1+a_Ku) + \ln u$$

The final indefinite integral I, after undoing substitution, is

$$= -\sigma \left\{ \frac{\frac{t}{\sigma}}{\left(1 + a_1 e^{\frac{t}{\sigma}}\right)\left(1 + a_2 e^{\frac{t}{\sigma}}\right) \dots \left(1 + a_K e^{\frac{t}{\sigma}}\right)} + \frac{a_1^{K-1}}{(a_1 - a_2)(a_1 - a_3)..(a_1 - a_K)} \ln\left(1 + a_1 e^{\frac{t}{\sigma}}\right) + \cdots + \frac{a_K^{K-1}}{(a_K - a_1)(a_K - a_2)..(a_K - a_{K-1})} \ln\left(1 + a_K e^{\frac{t}{\sigma}}\right) - \frac{t}{\sigma} \right\} + C$$

Putting limits of Integration:

$$As\ t \to -\infty,\quad \ln\left(1 + a_k e^{\frac{t}{\sigma}}\right) \to\ 0;\ and\ \left\{ \frac{t}{\left(1 + a_1 e^{\frac{t}{\sigma}}\right) \dots \left(1 + a_K e^{\frac{t}{\sigma}}\right)} - t \right\} \to 0\ [\mathbf{P1}]$$

$$\left(as \lim_{t \to -\infty} \frac{a_1 t e^{\frac{t}{\sigma}}}{1 + a_1 e^{\frac{t}{\sigma}}} = \lim_{t \to -\infty} \frac{a_1 t}{e^{-\frac{t}{\sigma}} + a_1} = \lim_{t \to -\infty} - \frac{a_1}{e^{-\frac{t}{\sigma}}} (Using\ L'Hospital) = 0\right.$$

Hence, as $t \to -\infty, I = 0$

As $t \to \infty$,

A property of partial fractions:

Equating the coefficient of the highest degree of u in $\dfrac{1}{u.(1 + a_1 u)(1 + a_2 u) \dots (1 + a_K u)}$

$$= \frac{1}{u} + \sum_{i=1}^{K} \frac{A_i}{1 + a_i u}\ ;\ where\ A_k = \frac{-a_k}{\prod_{j=1, j \neq k}^{j=K} (a_k - a_j)}$$

Coefficient of u^K *will be:* $\displaystyle\prod_{i=1}^{K} a_i + \sum_{i=1}^{K} (A_i \prod_{j=1, j \neq i}^{K} a_j) = 0$

Dividing both sides by $\displaystyle\prod_{i=1}^{K} a_i$*, we get:* $\displaystyle\sum_{i=1}^{K} \frac{A_i}{a_i} = -1 \to \sum_{k=1}^{K} \frac{a_k^{K-1}}{\prod_{j=1, j \neq k}^{j=K} (a_k - a_j)} = 1\ [\mathbf{P(2)}]$

Or, decomposing $\displaystyle\frac{t}{\sigma} = \sum_{k=1}^{K} \frac{\frac{t}{\sigma} a_k^{K-1}}{\prod_{j=1, j \neq k}^{j=K} (a_k - a_j)}$*, and simplifying each term:*

$$\lim_{t\to\infty}\ln\left(1+a_k e^{\frac{t}{\sigma}}\right)-\left(\frac{t}{\sigma}\right)=\lim_{t\to\infty}\ln\left(e^{\left(\ln\left(1+a_k e^{\frac{t}{\sigma}}\right)-\left(\frac{t}{\sigma}\right)\right)}\right)=\lim_{t\to\infty}\ln\left(e^{-\frac{t}{\sigma}}+a_k\right)=\ln a_k$$

$Hence, the\ definite\ integral\ becomes:$

$$\int_{-\infty}^{\infty} t.\frac{1}{\sigma}\frac{e^{\frac{t}{\sigma}}}{\prod_{k=1}^{K}\left(1+a_k e^{\frac{t}{\sigma}}\right)}\sum_{k=1}^{K}\left(\frac{a_k}{1+a_k e^{\frac{t}{\sigma}}}\right)=-\theta\sum_{k=1}^{K}\left\{a_k^{k-1}\frac{\ln a_k}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}\right\}$$

2. For the variance, we need to compute

$$E(\eta^2)=\int_{-\infty}^{\infty} t^2 f_\eta(t)dt=\int_{-\infty}^{\infty} t^2.\frac{1}{\sigma}\frac{e^{\frac{t}{\sigma}}}{\prod_{k=1}^{K}\left(1+a_k e^{\frac{t}{\sigma}}\right)}\sum_{k=1}^{K}\left(\frac{a_k}{1+a_k e^{\frac{t}{\sigma}}}\right)$$

Substituting $e^{\frac{t}{\sigma}}=u, t=\sigma\ln u, dt=\sigma.\frac{1}{u}du,$ and rewriting

$$\begin{aligned}I=\int_{-\infty}^{\infty}\frac{\sigma^2\ln^2 u}{\sigma}&\left\{\frac{u\ a_1\ (1+a_2u)(1+a_3u)\ldots(1+a_Ku)}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2}\right.\\&+\frac{u\ a_2\ (1+a_1u)(1+a_3u)\ldots(1+a_Ku)}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2}+\cdots\\&\left.+\frac{u\ a_K\ (1+a_1u)(1+a_3u)\ldots(1+a_{K-1}u)}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2}\right\}\sigma.\frac{1}{u}du\end{aligned}$$

$$=\sigma^2\int_0^{\infty}\ln^2 u\ .\frac{\begin{gathered}a_1\ (1+a_2u)(1+a_3u)\ldots(1+a_Ku)+\ a_2\ (1+a_1u)(1+a_3u)\ldots(1+a_Ku)+\cdots\\+a_K\ (1+a_1u)(1+a_3u)\ldots(1+a_{K-1}u)\end{gathered}}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2}$$

$Integrating\ by\ parts:$

$$Let\ I=\int fg'=fg-\int f'g$$

$$f=\ln^2 u\,,f'=2\frac{\ln u}{u}\ du$$

$$g'=\frac{\begin{gathered}a_1\ (1+a_2u)(1+a_3u)\ldots(1+a_Ku)+\ a_2\ (1+a_1u)(1+a_3u)\ldots(1+a_Ku)+\cdots\\+a_K\ (1+a_1u)(1+a_3u)\ldots(1+a_{K-1}u)\end{gathered}}{(1+a_1u)^2(1+a_2u)^2\ldots(1+a_Ku)^2}$$

$$g=-\frac{1}{(1+a_1u)(1+a_2u)\ldots(1+a_Ku)}$$

$$I=-\frac{\ln^2 u}{(1+a_1u)(1+a_2u)\ldots(1+a_Ku)}-\int -2\frac{\ln u}{u.(1+a_1u)(1+a_2u)\ldots(1+a_Ku)}du$$

$$Consider\ the\ Integral: J=\int\frac{\ln u}{u.(1+a_1u)(1+a_2u)\ldots(1+a_Ku)}du$$

Performing partial fractions decomposition:

$$J=-\frac{a_1^K}{(a_1-a_2)(a_1-a_3)..(a_1-a_K)}\int\frac{\ln u}{1+a_1u}du$$
$$-\frac{a_2^K}{(a_2-a_1)(a_2-a_3)..(a_2-a_K)}\int\frac{\ln u}{1+a_2u}du\ \dots$$
$$-\frac{a_K^K}{(a_K-a_1)(a_K-a_2)..(a_K-a_{K-1})}\int\frac{\ln u}{1+a_Ku}du+\int\frac{\ln u}{u}du$$

Consider the integral $M=\int\frac{\ln u}{1+a_ku}du$

Integrating by parts; let $f=\ln u\rightarrow f'=\frac{1}{u};\quad g'=\frac{1}{1+a_ku}\rightarrow\quad g=\frac{\ln(1+a_ku)}{a_k}$

$$\int fg'=fg-\int f'g\rightarrow M=\frac{\ln(u)\ln(1+a_ku)}{a_k}-\int\frac{\ln(1+a_ku)}{a_ku}du$$

Now Solving $N=\int\frac{\ln(1+a_ku)}{a_ku}$*; let* $z=-a_ku\rightarrow du=-\frac{dz}{a_k}$

$$N=-\frac{1}{a_k}\int-\frac{\ln(1-z)}{z}dz;\ \int-\frac{\ln(1-z)}{z}dz=Li_2(z)\rightarrow N=-\frac{1}{a_k}Li_2(z)$$

Undoing Substituion: $N==-\frac{Li_2(-a_ku)}{a_k}$ *and* $M=\frac{\ln(u)\ln(1+a_ku)}{a_k}-\frac{Li_2(-a_ku)}{a_k}$

Now solving $\int\frac{\ln u}{u}du$ *in* J*; Let* $\ln u=t;\frac{du}{u}=dt\rightarrow\int\frac{\ln u}{u}du=\int tdt=\frac{t^2}{2}=\frac{ln^2u}{2}$

$$Hence, J=\frac{\ln^2u}{2}-\frac{a_1^{K-1}}{(a_1-a_2)(a_1-a_3)..(a_1-a_K)}\{-\ln u\ln(1+a_1u)+Li_2(-a_1u)\}$$
$$-\frac{a_2^{K-1}}{(a_2-a_1)(a_2-a_3)..(a_2-a_K)}\{-\ln u\ln(1+a_2u)+Li_2(-a_2u)\}-\cdots$$
$$-\frac{a_K^{K-1}}{(a_K-a_1)(a_K-a_2)..(a_K-a_{K-1})}\{-\ln u\ln(1+a_Ku)+Li_2(-a_Ku)\}$$

Replacing for J, and rewriting the indefinite integral I after undoing substitution, we get:

$$I = -\sigma^2 \left\{ -\left(\frac{t}{\sigma}\right)^2 + \frac{\left(\frac{t}{\sigma}\right)^2}{\left(1+a_1e^{\frac{t}{\sigma}}\right)\left(1+a_2e^{\frac{t}{\sigma}}\right)\ldots\left(1+a_Ke^{\frac{t}{\sigma}}\right)} \right.$$

$$-2\left[\frac{a_1^{K-1}}{(a_1-a_2)(a_1-a_3)..(a_1-a_K)}\left\{-\frac{\mathrm{t}}{\theta}\ln\left(1+a_1e^{\frac{t}{\sigma}}\right)+Li_2\left(-a_1e^{\frac{t}{\sigma}}\right)\right\}\right.$$

$$+\frac{a_2^{K-1}}{(a_2-a_1)(a_2-a_3)..(a_2-a_K)}\left\{-\frac{\mathrm{t}}{\theta}\ln\left(1+a_2e^{\frac{t}{\sigma}}\right)+Li_2\left(-a_2e^{\frac{t}{\sigma}}\right)\right\}+\cdots$$

$$\left.\left.+\frac{a_K^{K-1}}{(a_K-a_1)(a_K-a_2)..(a_K-a_{K-1})}\left\{-\frac{\mathrm{t}}{\theta}\ln\left(1+a_Ke^{\frac{t}{\sigma}}\right)+Li_2\left(-a_Ke^{\frac{t}{\sigma}}\right)\right\}\right]\right\}$$

$$As\ t\to-\infty, Li_2\left(-a_1e^{\frac{t}{\sigma}}\right)=Li_2(0)=0,$$

$$and\ -\left(\frac{t}{\sigma}\right)^2+\frac{\left(\frac{t}{\sigma}\right)^2}{\left(1+a_1e^{\frac{t}{\sigma}}\right)\left(1+a_2e^{\frac{t}{\sigma}}\right)\ldots\left(1+a_Ke^{\frac{t}{\sigma}}\right)}=0\ \text{(By Property }\mathbf{P1}\text{)}$$

$$and\ \lim_{t\to-\infty}\frac{\mathrm{t}}{\theta}\ln\left(1+a_Ke^{\frac{t}{\sigma}}\right)=\lim_{t\to-\infty}\ln\left(e^{\frac{\mathrm{t}}{\theta}\ln\left(1+a_2e^{\frac{t}{\sigma}}\right)}\right)=e^{\frac{t}{\sigma}}\left(1+a_2e^{\frac{t}{\sigma}}\right)=0\ \Rightarrow I=0$$

$$At\ t=\infty\text{:}\ \lim_{t\to\infty}\frac{\left(\frac{t}{\sigma}\right)^2}{\left(1+a_1e^{\frac{t}{\sigma}}\right)\left(1+a_2e^{\frac{t}{\sigma}}\right)\ldots\left(1+a_Ke^{\frac{t}{\sigma}}\right)}=0$$

Also, We know: $Li_2(z)=-Li_2(z^{-1})-\frac{1}{2}\ln^2(-z)-\frac{\pi^2}{6}$

$$Li_2\left(-a_ke^{\frac{t}{\sigma}}\right)=\left(-Li_2\left(\frac{1}{-a_ke^{\frac{t}{\sigma}}}\right)-\frac{1}{2}\ln^2\left(a_ke^{\frac{t}{\sigma}}\right)-\frac{\pi^2}{6}\right)_{t\to\infty}=0-\frac{1}{2}\ln^2 a_k-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2-\frac{\pi^2}{6}$$

$$\lim_{t\to\infty}Li_2\left(-a_ke^{\frac{t}{\sigma}}\right)=\frac{-3\ln^2 a_k-\pi^2}{6}-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2$$

By Property $\mathbf{P(2)}$ $\sum_{k=1}^{K}\frac{a_k^{K-1}}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}=1\Rightarrow\sum_{k=1}^{K}\frac{a_k^{K-1}\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}=\frac{1}{2}\left(\frac{t}{\sigma}\right)^2$

Expanding, $I=-\sigma^2\{-\left(\frac{t}{\sigma}\right)^2-2\left[-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2+\sum_{k=1}^{K}\frac{a_k^{K-1}(-3\ln^2 a_k-\pi^2)}{6\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}\right]$

Simplifying : $I=E(\eta^2)=\sigma^2\sum_{k=1}^{K}\left(\frac{a_k^{K-1}(3\ln^2 a_k+\pi^2)}{3\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}\right),$

and $Var(\eta) = \sigma^2 \sum_{k=1}^{K} \left(\frac{a_k^{K-1}(3\ln^2 a_k + \pi^2)}{3\prod_{j=1, j\neq k}^{j=K}(a_k - a_j)} \right) - [E(\eta)]^2$

A.4: Expectation and Variance of Truncated MinLogistic Distribution

1. $E(\eta))|(\eta < c) = \int_{-\infty}^{c} t f_\eta(t)dt = I(t = c) - I(t = -\infty)$ (where I = Indefinite Integral)

$$I = -\sigma \left\{ \frac{\frac{t}{\sigma}}{\left(1 + a_1 e^{\frac{t}{\sigma}}\right)\left(1 + a_2 e^{\frac{t}{\sigma}}\right) \ldots \left(1 + a_K e^{\frac{t}{\sigma}}\right)} + \frac{a_1^{K-1}}{(a_1 - a_2)(a_1 - a_3)..(a_1 - a_K)} \ln\left(1 + a_1 e^{\frac{t}{\sigma}}\right) + \frac{a_2^{K-1}}{(a_2 - a_1)(a_2 - a_3)..(a_2 - a_K)} \ln\left(1 + a_2 e^{\frac{t}{\sigma}}\right) + \cdots + \frac{a_K^{K-1}}{(a_K - a_1)(a_K - a_2)..(a_K - a_{K-1})} \ln\left(1 + a_K e^{\frac{t}{\sigma}}\right) - \frac{t}{\sigma} \right\} + C$$

$I\ (t \to -\infty) = 0$ (found out earlier)

$$At\ t = c{:}\ I = -\sigma \sum_{k=1}^{K} \left\{ \frac{a_k^{K-1} \ln\left(1 + a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1, j\neq k}^{j=K}(a_k - a_j)} \right\} + c\left(1 - \frac{1}{\prod_{k=1}^{K}\left(1 + a_k e^{\frac{c}{\sigma}}\right)}\right)$$

$$So,\ \int_{-\infty}^{c} t f_\eta(t)dt = -\sigma \sum_{k=1}^{K} \left\{ \frac{a_k^{K-1} \ln\left(1 + a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1, j\neq k}^{j=K}(a_k - a_j)} \right\} + c\left(1 - \frac{1}{\prod_{k=1}^{K}\left(1 + a_k e^{\frac{c}{\sigma}}\right)}\right)$$

$$= c\int_{-\infty}^{c} f_\eta(t) - \sigma \sum_{k=1}^{K} \left\{ \frac{a_k^{K-1} \ln\left(1 + a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1, j\neq k}^{j=K}(a_k - a_j)} \right\}$$

$Normalizing\ the\ truncated\ expectation\ with\ \int_{-\infty}^{c} f_\eta(t)(= F_\eta(c)), We\ get{:}$

$$E(\eta))|(\eta < c) = c - \frac{1}{F_\eta(c)}\sigma \sum_{k=1}^{K} \left\{ \frac{a_k^{K-1} \ln\left(1 + a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1, j\neq k}^{j=K}(a_k - a_j)} \right\}$$

2. Variance of Truncated MinLogistic Distribution

$Consider\ \int_{-\infty}^{c} t^2 f_\eta(t)dt = I(t = c) - I(t \to -\infty)$ (where I is the indefinite Integral)

$I\ (t \to -\infty) = 0$ (found out earlier)

$$At\,t = c{:}\,I = -2\sigma^2\sum_{k=1}^{K}\frac{a_k^{K-1}Li_2\left(-a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - 2\sigma c\sum_{k=1}^{K}\frac{a_k^{K-1}ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} + c^2\left(1-\frac{1}{\prod_{k=1}^{K}\left(1+a_k e^{\frac{c}{\sigma}}\right)}\right)$$

$$= -2\sigma^2\sum_{k=1}^{K}\frac{a_k^{K-1}Li_2\left(-a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - 2\sigma c\sum_{k=1}^{K}\frac{a_k^{K-1}ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} + c^2F_\eta(c)$$

$After\ Normalisation{:}$

$$\int_{-\infty}^{c} t^2 f_\eta(t)dt = -\frac{2\sigma^2}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}Li_2\left(-a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - \frac{2\sigma c}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} + c^2$$

$$Var(\eta|\eta<c) = E(\eta^2|\eta<c) - [E(\eta|\eta<c)]^2$$

$$= -\frac{2\sigma^2}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}Li_2\left(-a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - \frac{2\sigma c}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} + c^2 - \left[c - \frac{1}{F_\eta(c)}\sigma\sum_{k=1}^{K}\left\{\frac{a_k^{K-1}\ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}\right\}\right]^2$$

$$= -\frac{2\sigma^2}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}Li_2\left(-a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - \frac{2\sigma c}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} + c^2 - c^2 + \frac{2\sigma c}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - \left[\frac{1}{F_\eta(c)}\sigma\sum_{k=1}^{K}\left\{\frac{a_k^{K-1}\ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}\right\}\right]^2$$

$$= -\frac{2\sigma^2}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}Li_2\left(-a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - \left[\frac{1}{F_\eta(c)}\sigma\sum_{k=1}^{K}\left\{\frac{a_k^{K-1}\ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}\right\}\right]^2$$

$$Var(\eta|\eta<c) = -\frac{2\sigma^2}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}Li_2\left(-a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - \left[\frac{1}{F_\eta(c)}\sigma\sum_{k=1}^{K}\left\{\frac{a_k^{K-1}\ln\left(1+a_k e^{\frac{c}{\sigma}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)}\right\}\right]^2$$

$$Var(\eta|\eta<c) = -\frac{2\sigma^2}{F_\eta(c)}\sum_{k=1}^{K}\frac{a_k^{K-1}Li_2\left(-a_k e^{\frac{c}{\theta}}\right)}{\prod_{j=1,j\neq k}^{j=K}(a_k-a_j)} - [c - E(\eta|\eta<c)]^2$$

APPENDIX B: Derivation of Forecasting Formula for Fractional Allocations

In this appendix, we derive the formula for a good *k*, conditional on the good being chosen for consumption. The allocation to the entire product group in determined in the first-level Tobit model (all fractional allocations are immediately zero if there is no allocation to the product group as a whole). So, consider the case of positive allocation to the product group, in which case at least one inside good should be consumed. Without any loss in generality, assume that the first good is consumed in such a situation. If only this inside good is consumed, the fractional prediction for this good should be one (which we will demonstrate later). But consider the more general case of multiple inside goods (say *M* goods, including the first good, and assume that these are the first *M* goods), with a zero fraction for the non-consumed inside goods. From the KKT conditions in Equations (10) and (11),

$$\tilde{\psi}_m\left(\frac{\tilde{f}_m^{op}}{\tilde{\gamma}_m}+1\right)^{-1}=\tilde{\psi}_1\left(\frac{\tilde{f}_1^{op}}{\tilde{\gamma}_1}+1\right)^{-1}\text{, because }\tilde{f}_m^{op}>0,\ \text{ or}$$
$$\tilde{f}_m^{op}=\left[\frac{\tilde{\psi}_m}{\tilde{\psi}_1}\left(\frac{\tilde{f}_1^{op}}{\tilde{\gamma}_1}+1\right)-1\right]\tilde{\gamma}_m,\ m=2,3,...,M. \qquad \text{(B.1)}$$

As in the text, the baseline preferences are denoted by $\tilde{\psi}_k$, and the satiation parameters by $\tilde{\gamma}_k$, because of the re-ordering of the goods from the highest value of ψ_k to the lowest. That is, $\tilde{\psi}_1>\tilde{\psi}_2>\tilde{\psi}_3>...\tilde{\psi}_K$. In particular, conditional on a positive allocation to the product group, the inside good with the highest baseline preference will definitely see some positive fractional allocation. Further, from the KKT conditions, the following should hold for the inside consumed goods 2 through *M*:

$$\tilde{\psi}_m>\tilde{\psi}_1\left(\frac{\tilde{f}_1^{op}}{\tilde{\gamma}_1}+1\right)^{-1}\ \text{ or }\ \left[\frac{\tilde{\psi}_m}{\tilde{\psi}_1}\left(\frac{\tilde{f}_1^{op}}{\tilde{\gamma}_1}+1\right)-1\right]>0\ \text{ for all }\ m=2,3,...,M, \qquad \text{(B.2)}$$

which immediately implies that $\tilde{f}_m^{op}>0$, *m*=2,3,…,*M*. Note also that the budget constraint $\sum_{j=1}^{M}\tilde{f}_j^{op}=1$ should be preserved by the optimal allocation formula for the fractions. For now, assume this is preserved (we will get back to showing this later, once the formula is derived). Then, it must be true that,

$$\tilde{f}_1^{op}=1-\sum_{j=2}^{M}\tilde{f}_j^{op}=1-\sum_{j=2}^{M}\left[\frac{\tilde{\psi}_j}{\tilde{\psi}_1}\left(\frac{\tilde{f}_1^{op}}{\tilde{\gamma}_1}+1\right)-1\right]\tilde{\gamma}_j. \qquad \text{(B.3)}$$

Solving for $\tilde{f}_1^{op}$, and after some re-arrangement, we get:

$$\tilde{f}_1^{op} = \frac{\tilde{\psi}_1\tilde{\gamma}_1 + \tilde{\gamma}_1 \sum_{j=2}^{M} \tilde{\gamma}_j \left(\tilde{\psi}_1 - \tilde{\psi}_j\right)}{\sum_{j=1}^{M} \tilde{\psi}_j \tilde{\gamma}_j}. \tag{B.4}$$

Note immediately that $\tilde{f}_1^{op}$>0, because $\tilde{\psi}_1 > \tilde{\psi}_j$ for all *j*=2,3,…,*K*. Substituting for $\tilde{f}_1^{op}$ back in the formula for $\tilde{f}_m^{op}$, and simplifying, we get the following:

$$\tilde{f}_m^{op} = \left[\frac{\tilde{\psi}_m}{\tilde{\psi}_1}\left(\frac{\tilde{f}_1^{op}}{\tilde{\gamma}_1} + 1\right) - 1\right]\tilde{\gamma}_m = \frac{\tilde{\psi}_m\tilde{\gamma}_m + \tilde{\gamma}_m \sum_{\substack{j=1 \\ j \neq m}}^{M} \tilde{\gamma}_j \left(\tilde{\psi}_m - \tilde{\psi}_j\right)}{\sum_{j=1}^{M} \tilde{\psi}_j \tilde{\gamma}_j}, \; m = 2,3,...,M. \tag{B.5}$$

Including good 1, we get the generic formula for any consumed good as in Equation (40) of the text:

$$\tilde{f}_m^{op} = \frac{\tilde{\psi}_m\tilde{\gamma}_m + \tilde{\gamma}_m \sum_{\substack{j=1 \\ j \neq m}}^{M} \tilde{\gamma}_j \left(\tilde{\psi}_m - \tilde{\psi}_j\right)}{\sum_{j=1}^{M} \tilde{\psi}_j \tilde{\gamma}_j}, \; m = 1,...,M \tag{B.6}$$

We have already shown that $\tilde{f}_m^{op}$>0, $m = 1,...,M$. It also is easy enough to show that $\sum_{m=1}^{M} \tilde{f}_m^{op} = 1$, because the second term, when summed across all inside goods M that are consumed is zero; that is, $\sum_{m=1}^{M} \tilde{\gamma}_m \sum_{\substack{j=1 \\ j \neq m}}^{M} \tilde{\gamma}_j \left(\tilde{\psi}_m - \tilde{\psi}_j\right) = 0.$ From the fact that $\tilde{f}_m^{op}$>0 for all consumed inside goods *m* and $\sum_{m=1}^{M} \tilde{f}_m^{op} = 1$, it immediately follows that $\tilde{f}_m^{op} \leq 1$ for each consumed inside good *m*. Of course, if only one inside good is chosen (the top good or good 1 in the descending order of baseline preference arrangement), $\tilde{f}_1^{op} = 1,$ as should be the case.

Next, note that, for any two consumed goods *l* and *m*, the KKT condition $\tilde{\psi}_m \left(\frac{\tilde{f}_m^{op}}{\tilde{\gamma}_m} + 1\right)^{-1} = \tilde{\psi}_l \left(\frac{\tilde{f}_l^{op}}{\tilde{\gamma}_l} + 1\right)^{-1}$ or $\left(\frac{\tilde{f}_m^{op}}{\tilde{\gamma}_m} + 1\right) / \left(\frac{\tilde{f}_l^{op}}{\tilde{\gamma}_l} + 1\right) = \tilde{\psi}_m / \tilde{\psi}_l$ should hold. This is guaranteed because $\left(\frac{\tilde{f}_m^{op}}{\tilde{\gamma}_m}+1\right)=\frac{\tilde{\psi}_m+\sum\limits_{\substack{j=1\\ j\neq m}}^{M}\tilde{\gamma}_j\left(\tilde{\psi}_m-\tilde{\psi}_j\right)}{\sum\limits_{j=1}^{M}\tilde{\psi}_j\tilde{\gamma}_j}+1=\frac{\tilde{\psi}_m\times\left(1+\sum\limits_{j=1}^{M}\tilde{\gamma}_j\right)}{\sum\limits_{j=1}^{M}\tilde{\psi}_j\tilde{\gamma}_j}$. Similarly,

$\left(\frac{\tilde{f}_l^{op}}{\tilde{\gamma}_m}+1\right)=\frac{\tilde{\psi}_l\times\left(1+\sum\limits_{j=1}^{M}\tilde{\gamma}_j\right)}{\sum\limits_{j=1}^{M}\tilde{\psi}_j\tilde{\gamma}_j}$, and the necessary equality results.

Finally, note that, for non-consumed goods, the KKT conditions of Equation (10) imply that the following should hold:

$$\psi_k<\psi_1\left(\frac{\tilde{f}_1^{op}}{\gamma_1}+1\right)^{-1},\ k=M+1,M+2,...,K.. \tag{B.7}$$

Substituting from Equation (B.4) for $\tilde{f}_1^{op}$, and after some algebra, we get:

$$\left(\frac{\tilde{f}_1^{op}}{\gamma_1}+1\right)=\frac{\psi_1+\psi_1\sum\limits_{j=1}^{M}\gamma_j}{\sum\limits_{j=1}^{M}\psi_j\gamma_j}. \tag{B.8}$$

Using (B.8) in (B.7), the result is the following:

$\psi_k-\left(\frac{\sum\limits_{j=1}^{M}\psi_j\gamma_j}{1+\sum\limits_{j=1}^{M}\gamma_j}\right)<0,\ k=M+1,M+2,...,K,$ or, equivalently,

$$\tilde{\pi}_k^{-1/\sigma}=\psi_k\gamma_k+\gamma_k\sum_{j=1}^{M}\gamma_j\left(\psi_k-\psi_j\right)<0,\ k=M+1,M+2,....,K.$$