Transformation-Based Flexible Error Structures for Choice Modeling

Chandra R. Bhat
The University of Texas at Austin
Dept of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712-1172
Phone: 512-471-4535, Fax: 512-475-8744
Email: bhat@mail.utexas.edu

ABSTRACT

In this paper, we propose a reverse Yeonjoo-Johnson (YJ) transformation to accommodate flexible skewed and fat-tailed specifications of stochastic terms in multivariate choice models. Essentially, we specify a YJ transformation of the univariate error terms to a univariate symmetric distribution, and then tie the resulting transformed univariate symmetric terms into a convenient symmetric multivariate distribution. In this paper, we use a normal distribution for the transformed univariate symmetric terms and bring these together using a multivariate normal distribution. In this way, the original non-normal error terms become reverse YJ-transformed. The use of such a flexible parametric distribution lends additional robustness to the maximum likelihood (ML) estimator. The proposed approach can be applied to a number of different univariate and multivariate mixed modeling choice structures. In a demonstration application, in the current paper, the proposed model is applied to investigate the effect of urban living on walking frequency, considering the choice of urban living as being endogenous to walking frequency.

Keywords: Multivariate choice, YJ transformation, Non-normality, Skewed distributions, Walking frequency.

1. INTRODUCTION

Econometric consumer choice models may have a variety of dependent outcome variables, including those that are continuous, grouped, binary, ordered-response, unordered-response (or nominal), count, or multiple- discrete-continuous. Increasingly, to recognize the jointness in decision-making regarding multiple dependent outcomes, there has been an emphasis on multivariate modeling approaches (see Bhat and Mondal, 2022 for a recent review of such approaches). In all these models, stochasticity is introduced to recognize that there will always be unobserved individual-specific factors that impact the outcome variable(s). Such stochasticity is commonly assumed to originate in the context of an underlying latent variable, which is then appropriately mapped to the actual observed outcome (except for the continuous case when the underlying latent variable is also the actual observed variable). It is typical to assume, purely out of convenience rather than based on any underlying economic theory, that a normal distribution characterizes such stochasticity, with the stochastic elements being brought together using a multivariate normal distribution for the case of joint multiple outcome analysis.

Two broad parametric methods are available in the literature to relax the *a priori* normal distribution assumption of stochastic terms, and generate flexible densities with skew and/or excess kurtosis (fat tails). The first is to use a parametric skew distribution (such as a skew-normal or a skew-t distribution or the broader class of skew-elliptical distributions or even mixtures of skew distributions) and the second is to use a transformation method. In its simplest version, the density in the former approach takes the form of the product of a symmetric density function and a skewing component that typically is the cumulative distribution of the symmetric density (see Lee and McLachlan, 2022). The latter transformation method essentially translates each marginal distribution into a univariate symmetric distribution (usually a normal distribution) and brings these together into a multivariate distribution across the marginals.[1] Of the two, the transformation method affords more flexibility as it is not tied to a specific density function, and can also mimic a whole host of skewed and fat-tailed density functions. As an illustration, Gallaugher et al. (2020) have compared, using Mardia's multivariate skewness and kurtosis metrics (Mardia, 1970) and multiple datasets for cluster analysis, the performance of two types of mixtures of skewed distributions with two transformation approaches. They conclude that "From the analyses on a variety of datasets…, it appears that no one method consistently outperforms the others and usually the performance is very similar if not identical". Further, as they state, the advantage of the transformation method is that it is much more parsimonious than the skewed approach, because the simple transformation approaches they use perform at least as well as complicated mixtures of skew distributions. The potency of the transformation method relative to the use of the more profligate density mixture approach is also evident in the extensive use of the transformation method in the field of data science and analytics as a preprocessing step in classification/regression type models to turn general distributions underlying the outcome variable into a near-normal

[1] Mixtures of univariate skew distributions for each marginal distribution (in the skew approach) or transforming the target density function to a mixture distribution (in the transformation approach) provide additional flexibility for each marginal. However, such mixtures entail additional parameters that get particularly difficult to estimate and implement for a high dimensional joint model with many outcome variables.

distribution (see, for example, Jadhav et al., 2023 and Peterson and Cavanaugh, 2020). Doing so not only improves the specification, but also benefits the power of statistical tests used in data analysis, decreasing the risk of committing Type 1 or Type II errors (Zimmerman, 1998 and Osborne, 2010). Thus, in this paper, we will focus on the transformation method for generating flexible distributions.

Of course, within the class of transformation methods, there are a whole suite of possible transformations to address skewness and fat tails, including square root transformation, inverse transformation, logarithmic transformation, arcsine transformation, Box-Cox (BC) transformation, and the Yeo-Johnson (YJ) transformation (see Box and Cox, 1964, Yeo and Johnson, 2000, Melnykov et al., 2021). Of these, the BC and YJ transformations have been shown, in analysis with both simulation and real empirical data, to be the best transformations in a wide variety of situations (see Osborne, 2010, Jadhav et al., 2023, Watthanacheewakul, 2021, and Marimuthu et al., 2022). In fact, most other transformations are special cases of the BC transformation for data on the positive half of the real line, depending on the value of the power parameter (for example, the square root transformation is the BC transformation with a power parameter of 0.5; the corresponding power parameters for the inverse and natural logarithm transformation are -1 and 0). Further, the YJ transformation is a simple extension of the restrictive BC transformation (which is applicable only to the positive half of the real line) to the entire real line, which makes it more appropriate for observable dependent outcomes that may take negative values too and to limited-dependent outcome models where the underlying latent variables span the entire real line. Thus, the YJ transformation allows for skew and fat tails, depending on the estimated value for an embedded transformation parameter. Bhat et al. (2024) discuss this YJ transformation in detail, demonstrating that it can mimic a whole variety of other flexible unimodal distributions closely, including the extreme value, the skew-normal, and skew-t distributions. Accordingly, in the current paper, we consider a YJ transformation for the unobserved term for the latent variable underlying limited dependent outcome models. We also extend the approach to multivariate limited-dependent outcome models, demonstrating an empirical application in the context of an ordered-response model with a binary endogenous explanatory variable or EEV. In this regard, the current paper constitutes the first formulation and application of a flexible multivariate non-normal limited-dependent variable model system with element-wise YJ transformation.

2. METHODOLOGY

We begin our discussion of the methodology with a simple univariate case of an ordered-response model (in our empirical analysis later, this outcome corresponds to walking frequency per week captured in five ordinal categories). However, the approach we propose is very general, and can be applied to literally any other limited-dependent outcome model, including unordered utility-based models, count models, grouped data models, and other econometric models. In Section 2.2, we show how to extend our approach to multivariate models, using a simple extension of the univariate ordered-response model as an example case.

2.1. Univariate Model Formulation

In the usual form of the ordered-response model structure (see McKelvey and Zavoina, 1971; Bhat, 1997; and Greene and Hensher, 2010), we write

$$y^* = \boldsymbol{\gamma}'\mathbf{x} + \delta u + \varepsilon, y = k \text{ if } \psi_{k-1} \le y^* < \psi_k; k = 1, 2, ...K; \ \psi_0 = -\infty, \psi_1 = 0, \psi_K = +\infty \quad (1)$$

In the above equation, the partitioning of the latent propensity y^* determines the ordered-response outcome y. $\mathbf{x}$ is a vector of exogenous variables (including a constant), while (for now) consider u as a binary (dummy) exogenous variable (say urban residence in the context of walk frequency).[2] $\boldsymbol{\gamma}$ is a coefficient vector to be estimated, and δ is another scalar parameter to be estimated.[3] In the usual way, the elements of the vectors $\mathbf{x}$ are assumed independent of the stochastic element ε. ψ_k is the upper bound threshold for ordinal level k for the outcome y ($\psi_0 < \psi_1 < \psi_2 ... < \psi_{K-1} < \psi_K$; $\psi_0 = -\infty, \psi_1 = 0, \psi_K = +\infty$). For future use, define $\boldsymbol{\psi} = (\psi_2, \psi_3, ..., \psi_{K-1})'$.

The canonical form of Equation (1) assumes that ε takes a standard logistic or standard normal distribution (the standardization is to account for scale unidentifiability). One way to generalize the error specification in Equation (1) is to accommodate asymmetry and skewness in ε. In earlier applications used in econometrics and machine learning, a strictly monotonic (and, therefore, one-to-one) transformation has been applied directly to y^* when y^* is observed as a continuous dependent outcome -- $\eta = t_\lambda(y^*)$, where η is closer to a symmetric distribution with non-fat tails (λ is a transformation parameter to be estimated). The transformed variable η is then used as the dependent outcome in the model (see Atkinson et al., 2021). One such restrictive transformation, of course, is the logarithmic transformation $\eta = t_\lambda(y^*) = \ln(y^*)$, with λ being an empty vector, in which case the untransformed outcome y^* is given by the inverse transformation $y^* = t_\lambda^{-1}(\eta) = \exp(\eta)$. Then, if η is assumed to be normally distributed, which is quite typical, y^* is log-normally distributed. However, much more flexible transformations, such as the YJ transformation may be used for $t_\lambda(.)$, while retaining the numerical convenience of the distribution of η. To see this, let the cumulative distribution function (CDF) of η be $F_\eta(z) = \text{Prob}(\eta < z)$ and let its probability density function (PDF) be $f_\eta(z)$. Given that the transformation $\eta = t_\lambda(y^*)$ is strictly monotonic, its inverse $y^* = t_\lambda^{-1}(\eta)$ exists, and the CDF of $y^* = t_\lambda^{-1}(\eta)$ is given by:

$$H_{y^*}(z) = \text{Prob}(y^* < z) = \text{Prob}(t_\lambda^{-1}(\eta) < z) = \text{Prob}(\eta < t_\lambda(z)) = F_\eta(t_\lambda(z)). \quad (2)$$

The corresponding PDF is given by:

[2] While u can be absorbed in the vector $\mathbf{x}$, separating it out will be helpful as we transition later to the specific application of our approach for an ordered-response model structure with an endogenous binary explanatory variable.

[3] The non-constant coefficients in $\boldsymbol{\gamma}$ can be considered as being random to accommodate unobserved heterogeneity in the sensitivity to the exogenous variables in $\mathbf{x}$. Doing so would not substantially change the basic formulation presented here, except that the resulting mixing would need appropriate integration. In this paper, we assume the non-constant coefficients in $\boldsymbol{\gamma}$ to be fixed to maintain focus on allowing a flexible distribution for the kernel error term ε, the main emphasis of the current paper.

$$h_{y^*}(z) = f_\eta\left(t_\lambda(z)\right)\left|\frac{\partial t_\lambda(z)}{\partial z}\right|. \tag{3}$$

Thus, the CDF of y^* takes the same convenient form as the CDF of η, except that the CDF of η is computed at a different point $t_\lambda(z)$ rather than at *z*. The PDF of y^* is also readily computed from the PDF of η computed at $t_\lambda(z)$.

In our case of Equation (1), however, the underlying continuous random variable y^* is not observed, and so we propose placing a transformation on the conditional (on $\mathbf{x}$ and u) underlying latent variable. That is, we apply the strictly monotonic transformation in reverse form by writing $\varepsilon = t_\lambda^{-1}(\eta)$, and maximize the resulting likelihood function to obtain an estimate of λ and other parameters of interest in the model. Defining $\varphi_k = \psi_k - \boldsymbol{\gamma}'\mathbf{x} - \delta u,$ the required probability corresponding to Equation (1) for maximum likelihood estimation is:

$$\text{Prob}(y = k) = \text{Prob}(\varphi_{k-1} \le \varepsilon < \varphi_k) = F_\eta\left[t_\lambda(\varphi_k)\right] - F_\eta\left[t_\lambda(\varphi_{k-1})\right] \tag{4}$$

Importantly, the convenience of the easily computed CDF of η is retained, while allowing a new, more flexible distribution for ε based on the transformation $t_\lambda(.)$. The point then is to consider a suitable transformation $t_\lambda(.)$ that allows for a flexible distribution of the error term ε that accommodates a range of asymmetric, skewed, and fat-tailed distributions. For example, in the empirical context of this paper, which focuses on walk frequency measured on an ordinal scale, it is highly unlikely that the unobserved term ε would have a symmetric distribution. Rather, we would expect that, given observed characteristics, a few individuals will have a high and spread-out intensity in their walking. But there is likely to be clustering at the lower end of walking intensity, especially given that there is also a bounding at zero. This would lead to a situation where there is a long right tail in the distribution of ε . More generally in empirical applications, there could be a variety of reasons why ε may not be symmetrically distributed. In the current paper, we propose the use of the YJ-transformation for the unobserved error term, such that ε is transformed into a symmetric distribution of η as follows:

$$\eta = t_\lambda(\varepsilon) = \begin{cases} -\dfrac{(-\varepsilon+1)^{2-\lambda}-1}{2-\lambda} & \text{if } \varepsilon < 0 \\ \dfrac{(\varepsilon+1)^{\lambda}-1}{\lambda} & \text{if } \varepsilon > 0 \end{cases} \tag{5}$$

However, we apply the above transformation in reverse to generate asymmetry and skew from a sample of symmetric data points drawn from the distribution of η as follows:

$$\varepsilon = t_\lambda^{-1}(\eta) = \begin{cases} 1-\left[1-(2-\lambda)\eta\right]^{\left(\frac{1}{2-\lambda}\right)} & \text{if } \eta < 0 \\ \left[1+\eta\lambda\right]^{\left(\frac{1}{\lambda}\right)} - 1 & \text{if } \eta > 0, \end{cases} \tag{6}$$

where $0 < \lambda < 2$. When $0 < \lambda < 1$, any value from the symmetric distribution of η on the real line is moved toward the right by the reverse transformation above (but without changing the sign of the

value of η). However, negative values for η are moved toward the right less than positive values for η, which results in a clustering closer to zero for the negative values of η and higher spread of η for the positive values of η. The net result is that ε gets skewed to the right with a long right tail. If $1<\lambda<2$, the opposite occurs and ε is skewed to the left with a long left tail. The original distribution of η is returned for ε if $\lambda=1$.

An additional issue is the distribution assumed for η. While any convenient symmetric distribution may be employed, we assume a univariate normal distribution, because of the flexibility to extend our reverse transformation technique to estimate general multivariate model systems with non-normal error terms, thanks to the convenient properties of the multivariate normal distribution. That is, we assume $\eta \sim N(\mu,\sigma^2)$. With this, the CDF and PDF of ε may be easily derived using the generic form of Equations (2) and (3) as:

$$H_\varepsilon(z) = \text{Prob}(t_\lambda^{-1}(\eta) < z) = \text{Prob}(\eta < t_\lambda(z)) = \Phi\left[\sigma^{-1}(t_\lambda(z)-\mu)\right]. \tag{7}$$

$$h_\varepsilon(z) = \frac{\partial H_\varepsilon(z)}{\partial z} = \frac{\partial \Phi\left[\sigma^{-1}(t_\lambda(z)-\mu)\right]}{\partial t_\lambda(z)} \times \left|\frac{\partial t_\lambda(z)}{\partial z}\right| = \left(\frac{\phi\left[\sigma^{-1}(t_\lambda(z)-\mu)\right]}{\sigma}\right) \times \left(|z|+1\right)^{\text{sgn}(z)(\lambda-1)}. \tag{8}$$

$\Phi(.)$ and $\phi(.)$ in the Equation above refer to the CDF and PDF of the standard univariate normal distribution, and $\text{sgn}(z)$ take the value of 1 if z is positive, the value of -1 if z is negative, and the value of 0 if z is zero. Figure 1 plots the PDF of ε for $\mu=0$ and $\sigma^2=1$, for different values of λ. The plots are restricted to the $0<\lambda\leq 1$ range, because the corresponding plots for $1\leq\lambda<2$ are mirror images of the plots shown, except with the skew toward the left. The plots show the flexibility of the YJ transformation to accommodate different levels of skew and tail thickness.

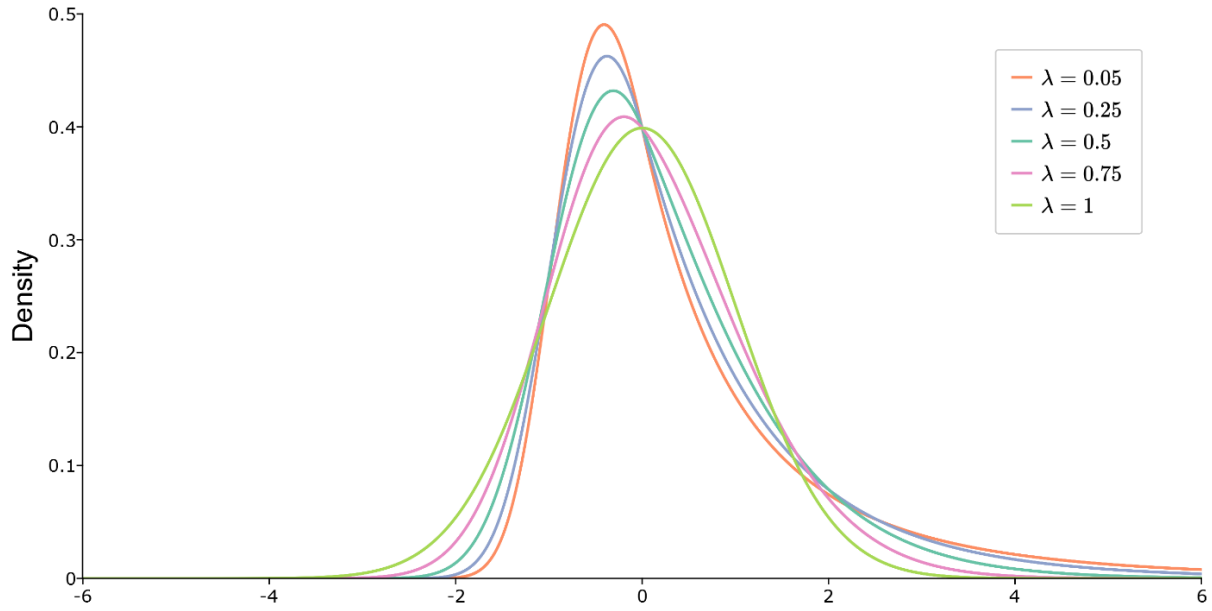

Figure 1: Density of transformed variable for different lambda values

Another important point here is that, for our case of an ordered-response model, we normalize the transformed variable η to be standard univariate normally distributed; that is, as in the plots of Figure 1, we assume $\mu = 0$ and $\sigma^2 = 1$ in model estimation. Of course, this does not imply that the error term ε will be normalized to a mean of zero and standard deviation of 1. The mean and standard deviation of ε will be determined by the YJ transformation parameter λ. In effect, λ is estimated so as to characterize the distribution of the ε (including the mean, the standard deviation, the skew, and tail thickness) that provides the best data fit to the observed discrete outcomes. This approach is compatible with the location and scale invariance of the latent variable y^*. Besides doing so also nests the univariate probit model (with the usual standard normalization on the error term) as a special case of the proposed model when λ=1.[4]

2.2. Multivariate Formulation

The reverse YJ transformation approach of the previous section can be extended to the case of multivariate outcomes. To do so, we include a subscript *l* for a set of *L* error terms stacked into a vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_L)'$ ($L \times 1$ vector). We then generate skew and asymmetry for each of the error terms using the reverse YJ transformation. Across different random variables $\varepsilon_1, \varepsilon_2, ..., \varepsilon_L$, the direction and intensity of skew/tail can vary. Then, the different variables ε_l $(l = 1, 2, ..., L)$ can be brought together into a multivariate distribution (see Loaiza-Maya et al., 2022 and Bhat et al., 2022). For convenience, define $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_L)'$ ($L \times 1$ vector), and $\mathbf{t}_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\eta}) = \left[t_{\lambda_1}^{-1}(\eta_1), t_{\lambda_2}^{-1}(\eta_2), ..., t_{\lambda_L}^{-1}(\eta_L) \right]$ ($L \times 1$ vector). Then, assuming a multivariate normal distribution for $\boldsymbol{\eta}$, we may write the cumulative distribution function for the vector $\boldsymbol{\varepsilon}$ as follows:

$$
\begin{aligned}
H_{\boldsymbol{\varepsilon}}(\mathbf{z}) &= \text{Prob}\left[\varepsilon_1 < z_1, \varepsilon_2 < z_2, ..., \varepsilon_L < z_L\right] = \text{Prob}(\boldsymbol{\varepsilon} < \mathbf{z}) \\
&= \text{Prob}\left[\mathbf{t}_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\eta}) < \mathbf{z}\right] = \text{Prob}\left[\boldsymbol{\eta} < \mathbf{t}_{\boldsymbol{\lambda}}(\mathbf{z})\right] \\
&= \Phi_L\left[\boldsymbol{\omega}^{-1}(\mathbf{t}_{\boldsymbol{\lambda}}(\mathbf{z}) - \boldsymbol{\mu}), \boldsymbol{\Omega}^*\right] \text{ with } \boldsymbol{\Omega}^* = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}, \\
\boldsymbol{\mu} &= (\mu_1, \mu_2, ..., \mu_L)', \ \boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1L} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1L} & \sigma_{2L} & \cdots & \sigma_L^2 \end{bmatrix},
\end{aligned}
\tag{9}
$$

where $\boldsymbol{\omega}$ is the diagonal matrix containing the square root of the variance elements of $\boldsymbol{\Omega}$, and Φ_L (.) in the equation above is the multivariate standard normal CDF of dimension L. The net result is that we have now generated asymmetry and skew for the multivariate distribution of $\boldsymbol{\varepsilon}$ through

[4] There is no known closed form for the moments of the untransformed error vector ε as a function of the YJ parameter vector λ when η is assumed to be normally distributed. But these moments can be estimated through numerical integration based on the probability density function in Equation (8).

the use of the element wise reverse YJ-transformation parameters λ_l from the multivariate normal distribution of $\boldsymbol{\eta}$. For completeness, we also write the corresponding multivariate density function as follows:

$$\begin{aligned} h_{\boldsymbol{\varepsilon}}(\mathbf{z}) &= \frac{\partial H_{\boldsymbol{\varepsilon}}(\mathbf{z})}{\partial \mathbf{z}'} = \frac{\partial \Phi_L\left[\boldsymbol{\omega}^{-1}(\mathbf{t}_{\boldsymbol{\lambda}}(\mathbf{z})-\boldsymbol{\mu}), \boldsymbol{\Omega}^*\right]}{\partial \mathbf{t}_{\boldsymbol{\lambda}}(\mathbf{z})} \times \left|\frac{\partial \mathbf{t}_{\boldsymbol{\lambda}}(\mathbf{z})}{\partial \mathbf{z}'}\right| \\ &= \left(\frac{\phi_L\left[\boldsymbol{\omega}^{-1}(\mathbf{t}_{\boldsymbol{\lambda}}(\mathbf{z})-\boldsymbol{\mu}), \boldsymbol{\Omega}^*\right]}{\prod_{l=1}^{L}\omega_l}\right) \times \prod_{l=1}^{L}\left(|z_l|+1\right)^{\text{sgn}(z_l)(\lambda_l-1)}, \end{aligned} \tag{10}$$

$\phi_L(.)$ is the multivariate standard normal pdf.

We now discuss the application of the multivariate transformation approach above in the specific context of an ordered-response model structure with an endogenous explanatory variable (EEV). In the case of Equation (1), for example, the dummy variable *u* may be endogenous to *y*. In our empirical analysis later, where *y* is walking frequency and *u* is a dummy variable for urban (versus) rural residence representing the built environment effect of residential location, a likely situation is the presence of unobserved antecedent personality, attitude, and lifestyle characteristics of individuals/households that simultaneously impact residential location choice and walk frequency. Bhat and Guo (2007) discuss this residential self-selection issue at length, emphasizing its importance for policy-making. To accommodate for this self-selection, while also recognizing the potentially non-normal nature of the unobserved factors affecting urban residence and walking frequency, we write two equations, one for the urban EEV and the other for the ordered-response:

$$\begin{aligned} &u^* = \boldsymbol{\beta}'\mathbf{w} + \varepsilon_1,\ u = 1 \text{ if } u^* \geq 0; u = 0 \text{ if } u^* < 0 \\ &y^* = \boldsymbol{\gamma}'\mathbf{x} + \delta u + \varepsilon_2,\ y = k \text{ if } \psi_{k-1} \leq y^* < \psi_k; k = 1, 2, ...K;\ \psi_0 = -\infty, \psi_1 = 0,\ \psi_K = +\infty \end{aligned} \tag{11}$$

The latent propensity u^* determines the EEV outcome u, with exogenous covariate vector $\mathbf{w}$ (including a constant). $\boldsymbol{\beta}$ is an additional coefficient vector to be estimated, and δ is the treatment effect parameter. The elements of the vectors $\mathbf{w}$ and $\mathbf{x}$ are assumed independent of the stochastic elements ε_1 and ε_2, but ε_1 and ε_2 are correlated. For identification, we also maintain the usual exclusion restriction that there is at least one variable ("instrument") that is contained in the vector $\mathbf{w}$, but does not appear in the vector $\mathbf{x}$. While there has been some confusion in the literature on whether such an exclusion restriction is necessary in multivariate limited-dependent variable models of the type in Equation (11) (see, for example, Wilde, 2000 and Rhine et al., 2006), Han and Lee (2019) show that such an exclusion restriction will, in general, be necessary and sufficient for identification of the model parameters. While the parameters may be identified even without the exclusion restriction in the specific case of the presence of a common continuous exogenous variable (but not a binary variable) in both $\mathbf{w}$ and $\mathbf{x}$ for a bivariate normal distribution of ε_1 and ε_2, only weak identification is suggested (at the very best) even in this case. And for more flexible

marginal distributions for ε_1 and ε_2, as in the current paper, the exclusion condition is necessary. Also, in the case when flexible marginal distributions for ε_1 and ε_2 are used (either in parametric or non-parametric form), the findings from Han and Lee (2019) hold that, for point identification of the parameters, a restrictive dependence structure has to be specified for the joint distribution of ε_1 and ε_2, even after allowing for an exclusion restriction. This restrictive dependence structure is satisfied by our use of the bivariate normal distribution (Gaussian copula) for the transformed (normal) marginals of ε_1 and ε_2, as discussed below.[5]

Next, write ε_1 and ε_2 in terms of their YJ-transformed normally distributed counterparts η_1 and η_2, and then construct a bivariate normal distribution for η_1 and η_2 as a special case of Equation (9):

$$\varepsilon_1 = t_{\lambda_1}^{-1}(\eta_1),\ \eta_1 \sim N(0,1); \varepsilon_2 = t_{\lambda_2}^{-1}(\eta_2),\ \eta_2 \sim N(0,1), \text{ and}$$
$$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} \sim BVN(\mathbf{0}, \boldsymbol{\Omega}^*), \text{ where } \mathbf{0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \boldsymbol{\Omega}^* = \begin{bmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \end{bmatrix}. \tag{12}$$

The specification in Equation (12) nests the bivariate probit model (with the usual standard normalization on the error terms) as a special case of the proposed model when $\lambda_1 = \lambda_2 = 1$.

To estimate the model, from Equation (9), we may write the cumulative distribution of ε_1 and ε_2 as follows:

$$H_{\varepsilon_1,\varepsilon_2}(z_1, z_2) = H_{\varepsilon_1,\varepsilon_2}(\mathbf{z}) = F_2\left[\mathbf{t}_{\boldsymbol{\lambda}}(\mathbf{z}); \mathbf{0}, \boldsymbol{\Omega}^*\right] = \Phi_2\left[\mathbf{t}_{\boldsymbol{\lambda}}(\mathbf{z}), \boldsymbol{\Omega}^*\right] = \Phi_2(t_{\lambda_1}(z_1), t_{\lambda_2}(z_2), \rho), \tag{13}$$

where $\Phi_2(.,.,\rho)$ refers to the standardized bivariate normal distribution with correlation ρ. Next, define $\theta = -\boldsymbol{\beta}'\boldsymbol{x}_1$, $\varphi_{k,1} = \psi_k - \boldsymbol{\gamma}'\boldsymbol{x} - \delta$, and $\varphi_{k,0} = \psi_k - \boldsymbol{\gamma}'\boldsymbol{x}$. The joint probabilities for the two regimes of $u = 0$ and $u = 1$ are:

$$\begin{aligned} P(u = 0, y = k) &= P(u^* < 0,\ \psi_{k-1} < y^* < \psi_k) = P(\varepsilon_1 < \theta, \varphi_{k-1,0} < \varepsilon_2 < \varphi_{k,0}) \\ &= H_{\varepsilon_1,\varepsilon_2}(\theta, \varphi_{k,0}) - H_{\varepsilon_1,\varepsilon_2}(\theta, \varphi_{k-1,0}) \\ &= \Phi_2\left(t_{\lambda_1}(\theta), t_{\lambda_2}(\varphi_{k,0}), \rho\right) - \Phi_2\left(t_{\lambda_1}(\theta), t_{\lambda_2}(\varphi_{k-1,0}), \rho\right), \text{ and} \end{aligned} \tag{14}$$

[5] These same identification issues also apply to joint models (such as a binary selection equation and an ordered-response equation) that do not have treatment parameters, since the likelihood function take identical forms with or without the treatment parameters in such cases. Also to be noted is that the control function (or two stage residual inclusion) method (see Rivers and Vuong 1988, Blundell and Powell 2004, Terza et al., 2008, Petrin and Train, 2010, and Wooldridge, 2015) to estimate models of the type in Equation system (11) is applicable only for continuous EEVs, and not for EEVs that have a limited range (including discrete EEVs) as in our case. The control function approach, as suggested by Terza et al. (2008), may work reasonably okay with limited-dependent EEVs when the endogeneity intensity is small (see Woolridge, 2015). But, as indicated by Wan et al. (2018), Mu and Zhang (2018) and Denzer (2019), such a control function approach requires the limited-dependent EEV to be a strictly increasing function of the first stage error. In other words, the first stage error (obtained as the difference between the limited-dependent EEV and its expected value in the population) is not independent of the exogenous variables in the EEV model, a necessary condition for the control function to be a consistent estimator. Thus, the control function is not, in general, appropriate for models with limited dependent EEVs.

$$
\begin{aligned}
P(u=1, y=k) &= P(u^* > 0,\ \psi_{k-1} < y^* < \psi_k) = P(\varepsilon_1 > \theta, \varphi_{k-1,1} < \varepsilon_2 < \varphi_{k,1}) \\
&= \left[H_{\varepsilon_2}(\varphi_{k,1}) - H_{\varepsilon_1,\varepsilon_2}(\theta, \varphi_{k,1}) \right] - \left[H_{\varepsilon_2}(\varphi_{k-1,1}) - H_{\varepsilon_1,\varepsilon_2}(\theta, \varphi_{k-1,1}) \right] \\
&= \left[\Phi\left(t_{\lambda_2}(\varphi_{k,1})\right) - \Phi_2\left(t_{\lambda_1}(\theta), t_{\lambda_2}(\varphi_{k,1}), \rho\right) \right] \\
&\quad - \left[\Phi\left(t_{\lambda_2}(\varphi_{k-1,1})\right) - \Phi_2\left(t_{\lambda_1}(\theta), t_{\lambda_2}(\varphi_{k-1,1}), \rho\right) \right] \\
&= \Phi_2\left(-t_{\lambda_1}(\theta), t_{\lambda_2}(\varphi_{k,1}), -\rho\right) - \Phi_2\left(-t_{\lambda_1}(\theta), t_{\lambda_2}(\varphi_{k-1,1}), -\rho\right)
\end{aligned} \tag{15}
$$

Now introduce the index q for individuals. Define a set of dummy variables M_{qk} (k=1,2,…,K): $M_{qk} = 1$ if $y = k$, and $M_{qk} = 0$ otherwise. Assuming independence across individuals, the likelihood function for estimation of the parameters in the model system is:

$$
L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \delta, \boldsymbol{\psi}, \rho) = \prod_{q=1}^{Q} \left(\left[P_q(u_q = 0, y_q = k) \right]^{(1-u_q)M_{qk}} \times \left[P_q(u_q = 1, y_q = k) \right]^{u_q M_{qk}} \right).^{6} \tag{16}
$$

In the maximization of the log-likelihood function, we need to ensure that the conditions $0 < (\lambda_1, \lambda_2) < 2$ strictly hold. For this, we use the following parameterization (showing this only for λ_1 with a similar parameterization for λ_2):

$$
\lambda_1 = \frac{2}{1 + \exp(-\lambda_1^*)}. \tag{17}
$$

Once the model is estimated with λ_1^* (and λ_2^*), we run a final iteration with the implied values of λ_1 and λ_2 to get the standard errors of all parameters, including λ_1 and λ_2.

3. APPLICATION TO URBAN RESIDENCE CHOICE AND WALKING FREQUENCY

3.1. Background

Walking is a travel mode that requires no motorization, is a natural human-learned skill from childhood for most individuals, entails zero monetary cost, can ease traffic congestion on our roadways, and contributes to a reduction of the transportation sector's carbon footprint (see Longo et al., 2015). At the individual level, walking, as a contributor to physical activity, can provide substantial health benefits, including both mental and physical.

An important issue that has attracted substantial interest in the land use-transportation literature in the past is the differential walking tendency based on residential built environment. Of keen interest over the years has been to disentangle associative effects from causal effects in

[6] The maximum likelihood (ML) method is predicated on the correct specification of the multivariate error structure underlying the overall limited dependent variable system (including the main outcome and the limited-dependent EEVs). Deviations from the correct distribution can lead to inconsistent estimation, especially in the effect of the limited-dependent EEV (sometimes referred to as the treatment effect when this happens to be a discrete EEV). This is where flexible parametric multivariate distributions, particularly the marginal distributions, can be substantially beneficial, while also being parsimonious in parameters and efficient in estimation. At the same time, because the ML estimator is a single-step estimator, standard errors are immediately available for all parameters. Further, the ML approach provides information on the joint (and marginal) distribution of the error terms.

the residential built environment (BE) influence on active travel behavior (for example, see Van Wee, 2009, Van Acker *et al.*, 2014, Bhat et al., 2016). This is particularly important when using cross-sectional data that entails the joint observation, at a specific point in time, of the residential location of households and activity-travel choices. From an econometric perspective, the "single-point-in-time" observation co-mingles the "true" BE effect of residential location with an associative effect due to the non-random assignment of individuals to residential locations. The question then is whether any co-movement between residential BE and walking intensity represents a true "causal" BE effect on walking intensity or simply an associative residential self-selection effect.

In the current paper, we focus on walking patterns among not-so-young adults (50+ age group), given the particularly large impact of COVID-19 on the active-travel (including walking) patterns of older adults. The BE attributes of the residential location of individuals in our study are captured using the single attribute of residential density, because of the well-established strong association between density and other BE elements (see, for example, Bhat and Singh, 2000, Ewing and Cervero, 2010, Kim and Brownstone, 2013, and Wang et al., 2021). As in most of these earlier studies, we use a binary classification of residential density as living in an urban area (associated with pedestrian-friendly BE) or a non-urban area (associated with a car-centric infrastructure environment) (this urban/non-urban classification was based on the metropolitan statistical area of residence).

The sample used in estimation is drawn from a walking survey of older adults in the US population used in the current paper was undertaken through the Foresight 50+ Consumer Omnibus panel survey, which is a mixed mode survey (online and telephone) funded and operated by the National Opinion Research Center (NORC) at the University of Chicago. The survey constitutes a probability-based panel designed to be representative of the US household population age 50 or older. The survey questions and survey instrument in the Foresight 50+ survey vary by month. The specific walking survey instrument used in the current paper was funded by the American Association of Retired Persons (AARP) and undertaken in July 2022.[7] A sample of 1691 individuals aged 50 years or older was obtained, of whom 1461 (86.4%) reside in urban areas. The ordered-response outcome in our empirical analysis corresponds to an ordinal scale of weekly walking frequency (in days per week of walking for at least 10 minutes) as follows: (a) Do not walk any day (19.2%), (b) Walk 1-2 days per week (17.0%), (c) Walk 3-4 days per week (27.9%), (d) Walk 5-6 days per week (21.1%), and (e) Walk 7 days per week (14.8%).

3.2. Model Results

A number of different specifications in terms of variables and the functional forms of variables were attempted in arriving at the final model specification. All variables, except age, are in either bracketed categories (income), or are naturally discrete. Examples of the latter include gender,

[7] Additional details of the survey administration and methodology are available at https://www.aarp.org/pri/topics/health/prevention-wellness/walking-attitudes-habits-adults-50-older/. Accessed November 13, 2023.

race/ethnicity, education level, housing tenure, dwelling type, employment status (paid employee, self-employed, unemployed/laid off temporarily, unemployed/looking for job, unemployed/retired, and unemployed/ physically disabled), household structure (single adult, single parent, married or unmarried couple, nuclear family, joint family of related individuals with more than two adults aged 18 or more and at least one individual of 17 years of age or younger, household with more than two adults aged 18 or more and no individual 17 years or younger), and residence region of the U.S.[8] The influence of the bracketed and discrete exogenous variables were tested as dummy variables in the most disaggregate form possible, and progressively combined for parsimony based on statistical tests. For the continuous age variable, alternative functional forms (such as a continuous linear form, a continuous logarithm form, a piece-wise linear form, and a set of dummy variables for different ranges) were tested. At the end, the effect of age was best represented in dummy variable form. Further, we examined interaction effects across variables, including between urban and gender, urban and race/ethnicity, and urban and household structure in the walk frequency equation (after accounting for the jointness through correlation effects between the urban residence and walk frequency endogenous outcomes), but none of these came out to be statistically significant.

The final model specification is presented in Table 1. The parameters for the urban or non-urban residence model component represent the elements of the $\boldsymbol{\beta}$ vector (that is, the effect of exogenous variables on the propensity to live in an urban region), while those for the walk frequency component refer to the elements of the $\boldsymbol{\gamma}$ vector and the urban "treatment" effect parameter δ (that is, the effect of exogenous variables and "urban" residence on the propensity underlying walking frequency). In Table 1, categories that do not appear in the column for urban living propensity or walk frequency propensity constitute the base categories. A '-' in the table implies that the corresponding coefficient was not statistically significant at even the 10% level of significance.

The results indicate that individuals from higher income households, with higher than a bachelor's degree, and older individuals (80 years or older) are more likely than their peers to be urban residents. In the context of a post-COVID landscape, these are quite interesting and suggest some shifts in residential choice patterns from before. There appears to be more of a move away from urban areas (to non-urban areas) among individuals with lower incomes and with low formal education degrees, in their quest to get better quality housing farther away, facilitated also by increased teleworking potential for erstwhile "blue collar" jobs (see Asmussen et al., 2024). Besides, recent studies have suggested that there has been a tempering in the market potential differential for employability between urban and non-urban areas, which would once again spur moves of low income individuals and those with low formal education degrees toward non-urban areas (Blumenberg and Wander, 2022). The higher tendency of older individuals to be located in

[8] The U.S. was geographically classified into four regions; Northeast, South, West, and Midwest; based on the Census Division of residence as follows: (a) New England and Mid-Atlantic = Northeast, (b) South Atlantic, West South Central and East South Central = South, (c) Pacific and Mountain=West, (d) East North Central and West North Central = Midwest.

urban areas may be tied to the reluctance to move away from current residences relative to younger individuals. Older individuals typically are averse to changes in life rhythms because stability provides a form of mental self-esteem boost (a sense of control and reduced stress/anxiety) for them at a time when their physical self-esteem may be on the decline (Duque et al., 2021).

Table 1: Main outcome results (coefficients represent effects on underlying latent propensity)

Variables	Urban Neighborhood Living		Walk Frequency (days per week)	
	Coeff.	t-stat	Coeff.	t-stat
Annual Household Income				
$75,000-$149,999	0.375	3.704		
$150,000 and higher	0.884	4.771		
$100,000 and higher	-	-	0.114	1.983
Higher than Bachelor's Degree	0.307	3.117	-	-
Woman	-	-	-0.124	-2.362
Age				
75 years or older	-	-	-0.231	-2.750
80 years or older	0.475	2.312	-	-
White Non-Hispanic	-0.736	-8.523	-	-
House Owned	-0.353	-2.953	-	-
Apartment Dwelling	0.523	3.390	-	-
Unemployed and Disabled	-	-	-0.397	-3.789
Unemployed and Retired	-	-	-0.124	-2.258
Geographic Residence Region				
South and Midwest	-0.358	-3.942	-	-
West	-	-	0.181	3.090
Correlation between the Normally Transformed Underlying Propensities between Urban Residence and Walking Frequency Error Terms	-0.257 (t-statistic -2.687)			
"True" Causal Effect of Urban Neighborhood Living	NA	NA	0.467	2.540
YJ Parameters (t-statistic computed with respect to the value of 1.000)	1.000*	-	0.526	2.843
Constant and Thresholds				
Constant	1.683	12.731	0.446	2.802
Threshold between walk 1/2 days per week and 3/4 days per week	-	-	0.440	1.934
Threshold between walk 3/4 days per week and 5/6 days per week	-	-	1.168	5.197
Threshold between walk 5/6 days per week and 7 days per week	-	-	2.068	9.053

* YJ parameter for the urban residence error term fixed to one because it was not statistically significantly different from one at any reasonable level of significance.

While some effects on residential location may have shifted in the aftermath of the COVID pandemic, others do not appear to have. In particular, individuals who identify as white non-Hispanic and those who live in an owned single-unit house (that is, do not live in rented apartment dwellings) are unlikely urban area residents. This is tied to non-white individuals having limited opportunities of residential mobility over decades, containing them to urban areas with a high supply of rental apartment dwellings. This may be traced to the long history of racial discrimination in the housing sector in the United States, offering more opportunities to white families to secure government-insured mortgages, while non-white families were not afforded the same level of benefits (Faber, 2020; Bhat et al., 2022). Also, individuals in the south and midwest of the country are less likely to be located in urban areas, as these parts of the country have large swaths of land with more spatial dispersion of residences.

In terms of walking frequency, individuals from high income households (100K per year or over) are more likely to walk over multiple days of the week relative to individuals from low income households (less than 100K per year), while women tend to have a lower walking propensity compared to men. Both of these results may be tied to time availability, with both low income earners and women well known to be time poor because of work-related and, in the case of women, also home-related responsibilities (see Bernardo et al., 2015, Cerrato and Cifre, 2018, and Mondal and Bhat, 2021). The result that older individuals (75 years and over), and those unemployed due to disabilities or who have retired, have a depressed walking propensity than their peers may be attributed to actual or perceived mobility limitations. From a geographic standpoint, residents from the west of the country are, in general, more predisposed to walk frequently.

As discussed earlier, we control for possible association in the residential choice-walking frequency choices when investigating the "true" causal effect of residential choice on walking frequency. Our results suggest that rural areas are better characterized (than urban areas) by unobserved basic-level walking needs of safety/comfort factors, which also promote walking frequency. This is evidenced by the negative correlation in the error terms between the urban area living propensity and the walk frequency equations. After accommodating for this association, the results show a clear positive "true" causal effect of urban living on walking frequency. If the association effects due to unobserved effects were ignored, the "true" urban effect gets underestimated. In fact, if the association is ignored, the urban effect becomes statistically insignificant at even the 65% confidence level (coefficient is 0.0740 with a t-statistic of 0.926). Overall, the results indicate that, in the aftermath of the worst of the pandemic, built environment factors associated with urban residence continue to have an important positive causal influence on walking activity engagement of "not-so-young" adults over the course of the week, consistent with earlier studies (see, for example, Lotfata et al., 2022). At the same time, the negative association between urban living and walk frequency suggests that walking environments need to be designed to feel less crowded (such as through the design of wide sidewalks and a well-distributed network of pedestrian walkways throughout the urban landscape to reduce pedestrian movement clustering), so that basic-level walking needs related to safe health environments are fulfilled.

The YJ parameters indicated that the urban equation error is not statistically significantly different from one at any reasonable level of significance, implying that we cannot reject that the urban error term is indeed normally distributed. So, we constrained the YJ parameter to the value of one. This normal marginal distribution for the urban living propensity is quite reasonable; there is no intuitive reason to expect skewness in this propensity. However, the YJ parameter for the walking frequency error term is significantly lower than one, indicating a rightward skew; that is, a small number of individuals have a high walk frequency, as we hypothesized in Section 2.1.

Finally, the constants and thresholds toward the end of the table do not have any substantive behavioral interpretation, though they serve the important purpose of mapping the latent propensities underlying the urban residence and walk frequency latent propensities to the corresponding categorical outcomes.

3.3. Data Fit Measures

We compare our proposed bivariate YJ model with five other restrictive versions: (i) a model that imposes error normality for both the urban residence and walk frequency outcomes as well as ignores endogeneity of urban residence (independent normal model), (ii) a model that considers the endogeneity in urban residence, but considers bivariate normality (bivariate normal model), (iii) a model that ignores the endogeneity in urban residence, but considers a skew-normal distribution for the walk frequency outcome (independent skew-normal model), (iv) a model that considers the endogeneity in urban residence and adopts a skew-normal distribution for the walk frequency outcome (bivariate skew-normal model), and (v) a model that ignores the endogeneity in urban residence, but considers a YJ-transformation for the walk frequency outcome (independent YJ model). As in our proposed bivariate YJ model, the hypothesis of normality of the error term in the urban residence outcome could not be ignored in the other non-normal model specifications too.

We consider the skew-normal distribution in our analysis as an empirical comparison point with our proposed approach. Conceptually, though, there are three reasons that one may prefer our YJ approach anyway in multivariate models. First, while there are different variants of the multivariate skew-normal distribution (see Arellano-Valle and Azzalini, 2006, and Lee and McLachlan, 2022), the most common one corresponds to skewness characterized by a single scalar variable, also referred to as the restricted multivariate skew normal (rMSN) distribution. This is the classical skew-normal distribution as proposed by Azzalini and Dalla Valle (1996). The advantage of the rMSN distribution is that the cumulative distribution function corresponds to that of a multivariate normal cdf with dimensionality one more than the dimensionality of the outcomes. This comes in handy when the number of outcomes in the model system is quite large (not much of an issue in our case because there are only two outcomes). The disadvantage is that independence cannot be maintained between any two error components within an rMSN distribution, except if one of them is normally distributed (that is, two univariate skew-normally distributed random terms that are tied together using the rMSN distribution will necessarily get correlated, thus not allowing for testing for independence; see Bhat and Sidharthan, 2012 for a

detailed discussion). Second, maximum likelihood estimation of skew-normal parameters can be difficult in many cases, especially limited-dependent model systems. This is because of the very close connection between the location, scale, and skew parameters in characterizing the skew-normal distribution. In particular, even when the location and scale parameters are set to zero and one, the mean is directly related to the skew parameter and so is the variance. Thus, especially when the location and scale of an underlying random variable is not identifiable, the skew parameter can be difficult to estimate, as we observed in our own estimations.[9] In our YJ approach, we do not have any such difficulty during estimation, because the transformation parameter λ is rather distinct from the parameters characterizing the distribution parameters of *G*. Third, the YJ transportation approach, coupled with a normal distribution for the YJ-transformed variables, is quite flexible and is able to mimic skew-normal distributions well (Bhat et al., 2024).

Turning back to the data fit from our empirical analysis, the adjusted likelihood ratio index of each model is first computed as follows with respect to the log-likelihood with only the constants/thresholds in the outcomes:

$$\bar{\rho}^2 = 1 - \frac{L(\hat{\boldsymbol{\theta}}) - M}{L(c)} \tag{18}$$

where $L(\hat{\boldsymbol{\theta}})$ is the log-likelihood function at convergence and $L(c)$ is the log-likelihood function with only the constants and the thresholds; *M* is the number of parameters (excluding the constants and thresholds) estimated in the model. Further, the "independent normal", the "bivariate model", and the "independent YJ" models are nested within our proposed "bivariate YJ model", making it easy to compare performances using the likelihood ratio test. The relevant likelihood-based data fit measures are provided in the top panel of Table 2. The $\bar{\rho}^2$ value for our proposed model is better than that for all the other models. Likelihood ratio (LR) tests (when the proposed model is compared to the three nested and restricted versions) yield values that are higher than the critical chi-squared table values at any reasonable significance level (at the respective degrees of freedom). The "independent skew-normal" and "bivariate skew-normal" are not nested within our proposed model, but can be compared using a non-nested likelihood ratio test. For example, let $\bar{\rho}^2_{\text{prop.YJ}}$ and $\bar{\rho}^2_{\text{ind.skew}}$ be the $\bar{\rho}^2$ for the proposed YJ and the independent skew models, respectively. If the difference in the indices is $(\bar{\rho}^2_{\text{prop.YJ}} - \bar{\rho}^2_{\text{ind.skew}}) = \tau$, then the probability that this difference could have occurred by chance is no larger than $\Phi\{-[-2\tau L(c) + (M_{GHDM} - M_{IROP})]^{0.5}\}$, 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.

[9]In estimating the skew-normal models, we first estimated the walk frequency ordered-response component separately. But the maximum likelihood estimation would not converge. So, we fixed the skew parameter at a battery of different values and estimated the model multiple times, and obtained the optimal skew value as the one at which the best (lowest) log-likelihood value was obtained. For the bivariate skew-normal model, we then fixed the skew parameter at what was estimated in the univariate walk frequency estimation and obtained an estimate of the correlation between the error terms of the urban residence and walk frequency equations, along with other parameter estimates.

The results for these non-nested LR tests are also provided in the top panel of Table 2, and again indicate the superior fit of our proposed YJ approach relative to the skew-normal approach.

In addition to the likelihood-based data fit measures, we also compute more intuitive non-likelihood based data fit metrics. At a disaggregate level, we compute the average (across individuals) probability of correct prediction (for the joint outcome of urban/non-urban living and walk frequency). At an aggregate level, we compare the predicted and actual (observed) numbers in each of the ten combinations of residence type and walk frequency, and compute a weighted average percentage error (WAPE) across all the ten possible combinations. The middle panel of Table 2 presents these non-likelihood data fit measures. The results clearly show that the predictions from our proposed model are closer to the observed values at both the disaggregate and aggregate levels. Indeed, the aggregate match of the predicted and actual values for our proposed bivariate YJ model is remarkable, especially given that the proposed model includes just one additional parameter (corresponding to the error term shape for the walk frequency model) relative to the bivariate normal model. Of course, we also will note that the WAPE for the other models are not too bad either, but the superior fit of the proposed model for each and every combination is clear. Interesting too is that, while the bivariate normal model (the bivariate skew-normal model) does have a statistically better log-likelihood at convergence at the 95% confidence level (and also a better average probability of correct prediction) relative to the independent normal model (the independent skew-normal model), the difference between these two models from a data fit perspective at the aggregate level is not too substantial (in fact, the independent normal and independent skew-normal models do marginally better than their respective bivariate counterparts; this can happen because the emphasis of the maximum likelihood estimation procedure is on maximizing the probability of the chosen alternative for each individual, not on aggregate predictions). In contrast, the proposed bivariate YJ model has a clear superior fit at both the disaggregate and aggregate levels.

Table 2: Data fit and ATE measures

Measure Type	Metric		Indep. Normal Model	Bivariate Normal Model	Indep. Skew-Normal Model	Bivariate Skew-Normal Model	Indep. YJ Model	Proposed Bivariate YJ Model
Likelihood Based Measures	Log-likelihood at convergence		-3215.54	-3213.46	-3212.24	-3209.13	-3211.29	-3207.49
	Number of non-constant/non-threshold parameters		15	16	16	17	16	17
	Log-likelihood at constants and thresholds only		-3353.17					
	Log-likelihood at zero (equal shares)		-3893.67					
	Rho-Bar Squared Value (w.r.t. constants/thresholds)		0.0366	0.0369	0.0369	0.0379	0.0375	0.0384
	LR test: Proposed Bivariate YJ vs Independent Normal		$LR = 16.1 > \chi^2_{(2,0.05)} = 5.99$					
	LR test: Proposed Bivariate YJ vs Bivariate Normal		$LR = 11.94 > \chi^2_{(1,0.05)} = 3.84$					
	LR test: Proposed Bivariate YJ vs Independent YJ		$LR = 7.6 > \chi^2_{(1,0.05)} = 3.84$					
	Non-Nested LR test: Proposed vs Independent Skew		Prob. that better fit of proposed model could have been random chance $= \Phi(-3.326) \approx 0$					
	Non-Nested LR test: Proposed vs Bivariate Skew		Prob. that better fit of proposed model could have been random chance $= \Phi(-1.831) \approx 0.035$					
Non-likelihood Based Measures	Average Probability of Correct Prediction (APCP)		0.1729	0.1731	0.1736	0.1740	0.1741	0.1748
	Actual versus Predicted Number of Individuals		Predicted					
	Combination Category	Actual						
	Non-urban residence and do not walk at all	57	48.48	48.34	50.79	50.89	53.04	56.16
	Non-urban residence/walk 1-2 days a week	36	41.11	41.51	41.13	42.00	40.89	37.41
	Non-urban residence/walk 3-4 days a week	56	64.35	64.93	63.18	63.81	61.44	55.63
	Non-urban residence/walk 5-6 days a week	45	45.90	46.04	45.04	44.91	44.39	44.61
	Non-urban residence/walk 7 days a week	36	30.26	30.04	29.96	29.53	30.33	36.82
	Urban residence and do not walk at all	268	276.62	276.76	274.60	274.63	272.84	269.62
	Urban residence/walk 1-2 days a week	252	248.64	248.74	247.71	247.48	246.90	251.05
	Urban residence/walk 3-4 days a week	416	407.94	407.43	408.57	407.88	410.00	415.94
	Urban residence/walk 5-6 days a week	311	307.77	306.99	310.01	309.43	311.63	310.63
	Urban residence/walk 7 days a week	214	219.93	220.22	220.00	220.43	219.54	213.13
	Weighted Average Percentage Error (WAPE)		3.42%	3.60%	2.95%	3.18%	2.53%	0.46%
Urban ATE Effect	ATE		0.062	0.812	0.117	1.080	0.163	1.004
	%ATE		1.8%	29.9%	3.5%	43.3%	5.0%	39.2%

3.4. Estimated Causal Effect of Urban Area Residence

The results from the previous section illustrate the better data fit of our proposed model, but does not provide an indication of the extent of mis-estimation (if at all) of the substantive "true" causal effect of urban residence on walk frequency. For this, we examine the impact of urban area residence using an average treatment effect or ATE (see Heckman and Vytlacil, 2000; Heckman and Vytlacil, 2001; and Bhat and Eluru, 2008).[10] In the current empirical context, the average treatment effect (ATE) refers to the expected walk frequency shift for a randomly picked household (from the entire pool of households independent of currently observed residential neighborhood) if it were to reside in an urban neighborhood relative to a non-urban neighborhood. It represents the "true" causal effect of urban residence on walk frequency. For ease in the computation of this ATE effect, we assign a cardinal value c_k to each ordinal walk frequency category k as follows: (a) Do not walk any day (k=1, $c_1 = 0$), (b) Walk 1-2 days per week (k=2, $c_2 = 1.5$), (c) Walk 3-4 days per week (k=3, $c_3 = 3.5$), (d) Walk 5-6 days per week (k=4, $c_4 = 5.5$), and (e) Walk 7 days per week (k=5, $c_5 = 7$). Then, the ATE metric is computed for our proposed model as follows (using the same notations as earlier):

$$\begin{aligned} ATE &= \frac{1}{Q}\sum_{q=1}^{Q}\left[\sum_{k=1}^{K} c_k \times\left[\Phi\left(t_{\lambda_2}(\varphi_{k,1})\right)-\Phi\left(t_{\lambda_2}(\varphi_{k-1,1})\right)\right]-\sum_{k=1}^{K} c_k \times\left[\Phi\left(t_{\lambda_2}(\varphi_{k,0})\right)-\Phi\left(t_{\lambda_2}(\varphi_{k-1,0})\right)\right]\right] \\ &= \frac{1}{Q}\sum_{q=1}^{Q}\left[\sum_{k=1}^{K} c_k \times\left\{\left[\Phi\left(t_{\lambda_2}(\varphi_{k,1})\right)-\Phi\left(t_{\lambda_2}(\varphi_{k-1,1})\right)\right]-\left[\Phi\left(t_{\lambda_2}(\varphi_{k,0})\right)-\Phi\left(t_{\lambda_2}(\varphi_{k-1,0})\right)\right]\right\}\right] \end{aligned} \tag{19}$$

For the more restrictive bivariate normal model, after estimation, we compute the ATE effect without the YJ transformations appearing in the equation above. For the independent YJ model, the ATE takes the same form as Equation (19) after estimation, while, for the independent normal model, the ATE takes the same form as for the bivariate normal model except that the coefficients correspond to the case when the error correlation is ignored. For the independent and bivariate skew-normal models, similar expressions (but using the probabilities based on the skew normal probability expressions) are computed. In addition, we also compute the %ATE effect, considering the walk frequency predicted by our proposed model for a random household if it were located in an urban location as the base. To be noted here is that we do not observe the walk frequency for a random household if were in an urban location relative to if it were in a non-urban location; these are counterfactual scenarios that can be obtained only through prediction. The counterfactual predictions are obtained from each model and used as the basis to compute the %ATE for the model (the counterfactual walk frequency prediction, in days of walk per week, for a random household if located in a non-urban area is 3.37 for the independent normal model; 2.72 for the

[10] The ATE effect, as discussed in this section to determine the urban neighborhood residence effect on walk frequency, can also be used to estimate the magnitude effect of any other sociodemographic, employment, or geographic residence region variable on walk frequency. However, we confine attention here to the urban residence effect.

bivariate normal model; 3.32 for the independent skew-normal model; 2.49 for the bivariate skew-normal model; 3.28 for the independent YJ model, and 2.55 for the bivariate YJ model).

The bottom panel of Table 2 presents these ATE effects for each of the six models. The first numeric value (0.062) indicates that a random household would walk an additional 0.062 days per week if in an urban setting rather than a non-urban setting, corresponding to a %ATE effect of 1.8% (that is, the independent normal model predicts that a random household would have 1.8% additional walk days if in an urban environment relative to a non-urban environment). The difference in ATE effects across the many models is rather clear. For the independent models, the ATE effect is substantially underestimated because of the negative correlation in the error terms of the urban residence equation and the walk frequency equation (that is, the fact that individuals with a high walk frequency propensity tend to locate themselves in non-urban areas gets comingled with the true "causal" effect, lowering the ATE substantially). The bivariate models disentangle the correlation effect from the "true" causal effect, which shows up as a much higher ATE effect. Even so, the bivariate normal model does underestimate the "true" causal effect (an ATE of 0.812 or about 30% ATE in the bivariate normal relative to an ATE of 1.004 or about 40% ATE in the bivariate YJ model). This underestimation of the "true" causal ATE effect of urban residence on walk frequency in the bivariate normal model is due to the asymmetry and rightward skew of the error term of the walk frequency propensity (as correctly recognized by the bivariate YJ model but ignored by the bivariate normal model). Of course, the extent of differences in the ATEs between a restrictive distribution model and the proposed more general distribution model will be specific to each empirical context considered, which implies that it is best to estimate the proposed model proposed rather than *a priori* settling for a restrictive structure that may provide inaccurate results. Also to be noted is that the ATE of the bivariate skew-normal is close to that of our proposed bivariate model, though slightly overestimated.

Finally, in terms of the residential self-selection effect, one may estimate this from the realization that any model that ignores the correlation between the urban residence and walk frequency equations comingles the associative effect of urban residence and the "true causal" effect of urban residence. Thus, considering the two YJ-based models, the ATE for the independent YJ model comingles the "spurious" self-selection and "true" causal urban residence effects, while the ATE for the bivariate model estimates the "true" causal effect. From the ATEs, one may then estimate the total of the two effects to be 1.004+(1.004-0.163)=1.845. Then, the self-selection effect amounts to 45.6% and the "true" causal effect amounts to 54.4%.

4. CONCLUSIONS

In this paper, we have shown the promise of the YJ transformation to accommodate flexible specifications of stochastic terms in multivariate mixed data models in general, and ordered-response models with discrete EEVs in particular. To our knowledge, this is the first such formulation and application in the econometric literature. The use of such a flexible parametric distribution leads to added robustness of the maximum likelihood (ML) estimator. The resulting multivariate YJ model is as easy to estimate as a multivariate normal model. More generally, the

YJ transformation is an efficient way to capture flexible marginal error distributions, which then can be bound together using a multivariate normal distribution. In the current paper, we have also shown that the proposed YJ approach is conceptually and empirically better than an a priori skew-normal specification for the error terms in an ordered-response model with an endogenous explanatory binary variable The proposed approach can be applied to a number of different univariate and multivariate mixed modeling structures, including sample selection models, endogenous switching models, multivariate mixed data models, and revealed preference-stated preference models. It can also be employed for empirical analysis in a variety of travel behavior, traffic safety, urban planning, education, public health, geography, and environmental economics fields, among other fields.

Of course, while providing an easy-to-estimate and efficient way to relax the typical multivariate normal distribution assumption used for the error terms in mixed data models, future research would benefit from an examination of the computational time/estimation stability gains from using our proposed approach, as well as comparing the potential flexibility losses of our proposed parametric approach, relative to the use of a multivariate nonparametric joint distribution for the error terms. Important to note, however, is that some assumptions related to smoothness and regularity need to be imposed even when adopting multivariate non-parametric joint distributions (see Gallant and Nychka, 1987, Vytlacil and Yildiz, 2007, or Chesher and Rosen 2013). Such nonparametric methods also quickly get extremely profligate in parameters because of series-based or similar approximations of the density function, especially when moving beyond a univariate distribution to a multivariate distribution (see Chen et al., 2006 and Denzer, 2019). Besides, no clear asymptotic distribution results are available for such models (needed to compute parameter standard errors), and any treatment effects of one endogenous outcome on another are not point-identified (Schwiebert, 2013; Han and Lee, 2019). Similar issues of profligateness and kernel density estimation difficulty can arise when using flexible semi-parametric approaches (such as those of Lewbel, 2000, Dong and Lewbel, 2015, Yildiz, 2013, and Mu and Zhang, 2018). These semi-parametric approaches also provide unbiased estimators only under specific conditions (such as a large support requirement for the "special regressor" in the Dong and Lewbel (2015) approach; see Bontemps and Nauges, 2017). In the above methodological context, what we have accomplished in this paper is to introduce our YJ-based formulation as a valuable addition to the toolbox of discrete choice modelers to introduce non-normality in mixed data models. But we leave, for future simulation-based work, extensive comparisons (from a stability, efficiency, efficacy, and computational standpoint) of the many possible non-normality approaches in mixed data models (with different numbers and types of endogenous outcomes, different sample sizes, different dependence patterns, and different univariate marginal error distributions).

From an empirical standpoint, the proposed model is applied in the current paper to investigate the effect of urban living on walking frequency, considering the choice of urban living as being endogenous to walking frequency. While this endogeneity has been attributed in the past to unobserved individual factors such as green consciousness, our results, in the context of a post-pandemic world, suggest that lower level needs of walking feasibility and safety related to health

and virus spread considerations may be the more dominant individual factors affecting walking frequency today and also leading to a generic predisposition of those with a high walking propensity to locate themselves in non-crowded non-urban areas. This is consistent with the findings from Paydar and Fard (2021) who note that high-density cities have been more vulnerable to the spread of COVID and so "more basic walking needs may play a more important role in the daily walking patterns of inhabitants". This is particularly so for the "not-so-young", because of the increased vulnerability of older individuals to COVID. The net result is a need to understand how best to design walking infrastructure that not only adheres to the traditional notions of pedestrian friendliness (such as narrow streets, continuity in walkways, frequent crossing points, and low speed limits), but also provides a sense of health safety from contagion spread. Doing so proactively would be beneficial in preparation for future pandemics. The takeaway also is that, with shifting residential choices, improved walking infrastructure (as proxied by the "urban" variable in our analysis) can elevate walking frequency across the board in all geographic areas, pointing to the importance of supporting policies to invest in walk-friendly environments in a post-pandemic world.

ACKNOWLEDGMENTS

This research was partially supported by the U.S. Department of Transportation through the Center for Understanding Future Travel Behavior and Demand (TBD) (Grant No. 69A3552344815 and No. 69A3552348320). The author is grateful to Lisa Macias for help in formatting this document. Two anonymous reviewers provided valuable feedback to improve the paper.

REFERENCES

Arellano-Valle, R.B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* 33(3), 561-574.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* 83(4), 715-726.

Asmussen, K.E., Mondal, A., and Bhat, C.R. (2024). The interplay between teleworking choice and commute distance. *Transportation Research Part C*, 165, 104690.

Atkinson, A.C., Riani, M., and Corbellini, A. (2021). The Box–Cox transformation: Review and extensions. *Statistical Science*, 36(2) 239-255. https://doi.org/10.1214/20-STS778

Bernardo, C., Paleti, R., Hoklas, M., and Bhat, C.R. (2015). An empirical investigation into the time-use and activity patterns of dual-earner couples with and without young children. *Transportation Research Part A*, 76, 71-91.

Bhat, C.R. (1997). Work travel mode choice and number of nonwork commute stops. *Transportation Research Part B*, 31(1), 41-54.

Bhat, C.R., and Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, 43(7), 749-765.

Bhat, C.R., and Guo, J.Y. (2007). A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B*, 41(5), 506-526.

Bhat, C.R., and Mondal, A. (2022). A new flexible generalized heterogeneous data model (GHDM) with an application to examine the effect of high density neighborhood living on bicycling frequency. *Transportation Research Part B*, 164, 244-266.

Bhat, C.R., and Sidharthan, R. (2012). A new approach to specify and estimate non-normally mixed multinomial probit models, *Transportation Research Part B*, 46, 817-833.

Bhat, C.R., and Singh, S.K. (2000). A comprehensive daily activity-travel generation model system for workers. *Transportation Research Part A*, 34(1), 1-22.

Bhat, C.R., Astroza, S., Bhat, A.C., and Nagel, K. (2016). Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: Application to residential self-selection effects analysis in an activity time-use behavior model. *Transportation Research Part B*, 91, 52-76.

Bhat, A.C., Almeida, D.M., Fenelon, A., Santos-Lozada, A.R. (2022). A longitudinal analysis of the relationship between housing insecurity and physical health among midlife and aging adults in the United States. *SSM - Population Health*, 18, 101128. https://doi.org/10.1016/j.ssmph.2022.101128

Bhat, C.R., Mondal, A., and Pinjari, A.R. (2024). A flexible non-normal random coefficient multinomial probit model: application to investigating commuter's mode choice behavior in a developing economy context. Technical paper, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin.

Blumenberg, E., and Wander, M. (2023). Housing affordability and commute distance. *Urban Geography*, 44(7), 1454-1473.

Blundell, R., and Powell, J. (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3), 655-679.

Bontemps, C., and Nauges, C. (2017). Endogenous variables in binary choice models: Some insights for practitioners. Working paper # 17-855, Toulouse School of Economics, https://publications.ut-capitole.fr/id/eprint/25726/1/wp_tse_855.pdf, accessed September 7, 2024.

Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), 211-252.

Cerrato, J., and Cifre, E. (2018). Gender inequality in household chores and work-family conflict. *Frontiers in Psychology*, 9, 1330.

Chen, X., Fan, Y., and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association*, 101(475), 1228-1240.

Chesher, A., and Rosen, A. (2013). What do instrumental variable models deliver with discrete dependent variables? *American Economic Review*, 103(3), 557-62.

Denzer, M. (2019). Estimating causal effects in binary response models with binary endogenous explanatory variables: A comparison of possible estimators. Discussion paper number 1916, Gutenberg School of Management and Economics, Johannes Gutenberg University Mainz, Germany.

Dong, Y., and Lewbel, A. (2015). A simple estimator for binary choice models with endogenous regressors. *Econometric Reviews*, 34(1-2), 82-105.

Duque, M. (2021). Performing healthy ageing through images: From broadcasting to silence. *Global Media and China*, 6(3), 303-324.

Ewing, R., and Cervero, R. (2010). Travel and the built environment. *Journal of the American Planning Association*, 76(3), 265-294.

Faber, J.W. (2020). We built this: Consequences of new deal era intervention in America's racial geography. *American Sociological Review*, 85(5), 739-775.

Gallant, A.R., and Nychka, D.W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55(2), 363-390.

Gallaugher, M.P.B, McNicholas, P.D., Melnykov, V., Zhu, X. (2020). Skewed distributions or transformations? Modelling skewness for a cluster analysis. https://doi.org/10.48550/arXiv.2011.09152, accessed January 7, 2024

Greene, W.H., and Hensher, D.A. (2010). *Modeling Ordered Choices: A Primer*. Cambridge University Press.

Han, S., and Lee, S. (2019). Estimation in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Applied Econometrics*, 34(6), 994-1015.

Heckman, J.J., and Vytlacil, E.J. (2000). The relationship between treatment parameters within a latent variable framework. *Economics Letters*, 66(1), 33-39.

Heckman, J.J., and Vytlacil, E.J. (2001). Policy-relevant treatment effects. *The American Economic Review*, 91, 107-111

Jadhav, A., Dhaulakhandi, D., Kumar, S., Malviya, L., and Mewada, A. (2023). Data transformation: A preprocessing stage in machine learning regression problems. In *Artificial Intelligence Techniques in Power Systems Operations and Analysis*, Singh, N., Tamrakar, S., Mewada, A., and Gupta, S.K. (Eds.). Auerbach Publications.

Kim, J. and Brownstone, D. (2013). The impact of residential density on vehicle usage and fuel consumption: Evidence from national samples. *Energy Economics*, 40, 196-206.

Lee, S.X., and McLachlan, G.L. (2022). An overview of skew distributions in model-based clustering. *Journal of Multivariate Analysis*, 188, 104853.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97(1), 145-177.

Lotfata, A., Gemci, A., and Ferah, B. (2022). The changing context of walking behavior: Coping with the COVID-19 pandemic in urban neighborhoods. *Archnet-IJAR: International Journal of Architectural Research*, 16(3), 495-516.

Longo, A., Hutchinson, W. G., Hunter, R. F., Tully, M. A., and Kee, F. (2015). Demand response to improved walking infrastructure: A study into the economics of walking and health behaviour change. *Social Science & Medicine*, 143, 107-116.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519-530.

Marimuthu, S., Mani, T., Sudarsanam, T.D., George, S., Jeyaseelan, L. (2022). Preferring Box-Cox transformation, instead of log transformation to convert skewed distribution of outcomes to normal in medical research. *Clinical Epidemiology and Global Health*, 15, 101043.

McKelvey, R.D., and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103-120.

Melnykov, Y., Zhu, X., and Melnykov, V. (2021). Transformation mixture modeling for skewed data groups with heavy tails and scatter. *Computational Statistics*, 36, 61-78.

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.

Mu, B., and Zhang, Z. (2018). Identification and estimation of heteroscedastic binary choice models with endogenous dummy regressors. *Econometrics Journal*, 21(2), 218-246.

Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, 15, Article 12.

Paydar, M., and Kamani Fard, A. (2021). The hierarchy of walking needs and the COVID-19 pandemic. *International Journal of Environmental Research and Public Health*, 18(14), 7461.

Peterson, R.A., and Cavanaugh, J.E. (2020). Ordered quantile normalization: A semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, 47(13-15), 2312-2327.

Petrin, A., and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47(1), 3-13.

Rhine, S.L.W., Greene, W.H., and Toussaint-Comeau, M. (2006). The importance of check-cashing businesses to the unbanked: Racial/ethnic differences. *The Review of Economics and Statistics*, 88(1), 146-157.

Rivers, D., and Vuong, Q.H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39(3), 347-366.

Schwiebert, J. (2013). Sieve maximum likelihood estimation of a copula-based sample selection model. Leibniz University Hannover, Institute of Labor Economics, https://conference.iza.org/conference_files/SUMS_2013/schwiebert_j8731.pdf, accessed September 7, 2024.

Terza, J.V., Basu, A., and Rathouz, P.J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3), 531-543.

Van Acker, V., Mokhtarian, P.L., and Witlox, F. (2014). Car availability explained by the structural relationships between lifestyles, residential location, and underlying residential and travel attitudes. *Transport Policy*, 35, 88-99.

Van Wee, B. (2009). Self-Selection: A key to a better understanding of location choices, travel behaviour and transport externalities? *Transport Reviews*, 29(3), 279-292.

Vytlacil, E., and Yildiz, N. (2007). Dummy endogenous variables in weakly separable models. *Econometrica*, 75(3), 757-779.

Wan, F., Small, D., and Mitra, N. (2018). A general approach to evaluating the bias of 2-stage instrumental variable estimators. *Statistics in Medicine*, 37(12), 1997-2015.

Wang, J., Yang, Y., Peng, J., Yang, L., Gou, Z., Lu, Y. (2021). Moderation effect of urban density on changes in physical activity during the coronavirus disease 2019 pandemic. *Sustainable Cities and Society*, 72, 103058.

Watthanacheewakul, L. (2021). Transformations for left skewed data. Proceedings of the World Congress on Engineering 2021 (WCE 2021), July 7-9, 2021, London, U.K.

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressors. *Economics Letters*, 69(3), 309-312.

Wooldridge, J.M. (2015). Control function methods in applied econometrics. *The Journal of Human Resources*, 50(2), 420-445.

Yeo, I.K., and Johnson, R.A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4), 954-959.

Yildiz, N. (2013). Estimation of binary choice models with linear index and dummy endogenous variables. *Econometric Theory*, 29(2), 354-92.

Zimmerman, D.W. (1998). Invalidation of parametric and nonparamteric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, 67(1), 55-68.