A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling

Chandra R. Bhat*
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
1 University Station C1761, Austin, TX 78712-0278
Phone: 512-471-4535, Fax: 512-475-8744
Email: bhat@mail.utexas.edu

and

Naveen Eluru
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
1 University Station, C1761, Austin, TX 78712-0278
Phone: 512-471-4535, Fax: 512-475-8744
Email: naveeneluru@mail.utexas.edu

*corresponding author

ABSTRACT

The dominant approach in the literature to dealing with sample selection is to assume a bivariate normality assumption directly on the error terms, or on transformed error terms, in the discrete and continuous equations. Such an assumption can be restrictive and inappropriate, since the implication is a linear and symmetrical dependency structure between the error terms. In this paper, we introduce and apply a flexible approach to sample selection in the context of built environment effects on travel behavior. The approach is based on the concept of a "copula", which is a multivariate functional form for the joint distribution of random variables derived purely from pre-specified parametric marginal distributions of each random variable. The copula concept has been recognized in the statistics field for several decades now, but it is only recently that it has been explicitly recognized and employed in the econometrics field. The copula-based approach retains a parametric specification for the bivariate dependency, but allows testing of several parametric structures to characterize the dependency. The empirical context in the current paper is a model of residential neighborhood choice and daily household vehicle miles of travel (VMT), using the 2000 San Francisco Bay Area Household Travel Survey (BATS). The sample selection hypothesis is that households select their residence locations based on their travel needs, which implies that observed VMT differences between households residing in neo-urbanist and conventional neighborhoods cannot be attributed entirely to the built environment variations between the two neighborhoods types. The results indicate that, in the empirical context of the current study, the VMT differences between households in different neighborhood types may be attributed to both built environment effects and residential self-selection effects. As importantly, the study indicates that use of a traditional Gaussian bivariate distribution to characterize the relationship in errors between residential choice and VMT can lead to misleading implications about built environment effects.

Keywords: copula; multivariate dependency; self-selection; treatment effects; vehicle miles of travel; maximum likelihood; archimedean copulas

1. INTRODUCTION

There has been considerable interest in the land use-transportation connection in the past decade, motivated by the possibility that land-use and urban form design policies can be used to control, manage, and shape individual traveler behavior and aggregate travel demand. A central issue in this regard is the debate whether any effect of the built environment on travel demand is causal or merely associative (or some combination of the two; see Bhat and Guo, 2007). To explicate this, consider a cross-sectional sample of households, some of whom live in a neo-urbanist neighborhood and others of whom live in a conventional neighborhood. A neo-urbanist neighborhood is one with high population density, high bicycle lane and roadway street density, good land-use mix, and good transit and non-motorized mode accessibility/facilities. A conventional neighborhood is one with relatively low population density, low bicycle lane and roadway street density, primarily single use residential land use, and auto-dependent urban design. Assume that the vehicle miles of travel (VMT) of households living in conventional neighborhoods is higher than the VMT of households residing in neo-urbanist neighborhoods. The question is whether this difference in VMT between households in conventional and neo-urbanist households is due to "true" effects of the built environment, or due to households self-selecting themselves into neighborhoods based on their VMT desires. For instance, it is at least possible (if not likely) that unobserved factors that increase the propensity or desire of a household to reside in a conventional neighborhood (such as overall auto inclination, a predisposition to enjoying travel, safety and security concerns regarding non-auto travel, *etc.*) also lead to the household putting more vehicle miles of travel on personal vehicles. If this self selection is not accounted for, the difference in VMT attributed directly to the variation in the built environment between conventional and neo-urbanist neighborhoods can be mis-estimated.

On the other hand, accommodating for such self-selection effects can aid in identifying the "true" causal effect of the built environment on VMT.

The situation just discussed can be cast in the form of Roy's (1951) endogenous switching model system (see Maddala, 1983; Chapter 9), which takes the following form:

$$\begin{aligned} & r_q^* = \beta' x_q + \varepsilon_q, \quad r_q = 1 \text{ if } r_q^* > 0, \quad r_q = 0 \text{ if } r_q^* \leq 0, \\ & m_{q0}^* = \alpha' z_q + \eta_q, \; m_{q0} = 1[r_q = 0] m_{q0}^* \\ & m_{q1}^* = \gamma' w_q + \xi_q, \; m_{q1} = 1[r_q = 1] m_{q1}^* \end{aligned} \tag{1}$$

The notation $1[r_q = 0]$ represents an indicator function taking the value 1 if $r_q = 0$ and 0 otherwise, while the notation $1[r_q = 1]$ represents an indicator function taking the value 1 if $r_q = 1$ and 0 otherwise. The first selection equation represents a binary discrete decision of households to reside in a neo-urbanist built environment neighborhood or a conventional built environment neighborhood. r_q^* in Equation (1) is the unobserved propensity to reside in a conventional neighborhood relative to a neo-urbanist neighborhood, which is a function of an (*M* x 1)-column vector x_q of household attributes (including a constant). β represents a corresponding (*M* x 1)-column vector of household attribute effects on the unobserved propensity to reside in a conventional neighborhood relative to a neo-urbanist neighborhood. In the usual structure of a binary choice model, the unobserved propensity r_q^* gets reflected in the actual observed choice r_q (r_q = 1 if the *q*th household chooses to reside in a conventional neighborhood, and r_q = 0 if the *q*th household decides to reside in a neo-urbanist neighborhood). ε_q is usually a standard normal or logistic error tem capturing the effects of unobserved factors on the residential choice decision.

The second and third equations of the system in Equation (1) represent the continuous outcome variables of log(vehicle miles of travel) in our empirical context. m_{q0}^* is a latent variable representing the logarithm of miles of travel if a random household q were to reside in a neo-urbanist neighborhood, and m_{q1}^{*} is the corresponding variable if the household q were to reside in a conventional neighborhood. These are related to vectors of household attributes z_q and w_q, respectively, in the usual linear regression fashion, with η_q and ξ_q being random error terms. Of course, we observe m_{q0}^{*} in the form of m_{q0} only if household q in the sample is observed to live in a neo-urbanist neighborhood. Similarly, we observe m_{q1}^{*} in the form of m_{q1} only if household q in the sample is observed to live in a conventional neighborhood.

The potential dependence between the error pairs (ε_q, η_q) and (ε_q, ξ_q) has to be expressly recognized in the above system, as discussed earlier from an intuitive standpoint.[1] The classic econometric estimation approach proceeds by using Heckman's or Lee's approaches or their variants (Heckman, 1974, 1976, 1979, 2001, Greene, 1981, Lee, 1982, 1983, Dubin and McFadden, 1984). Heckman's (1974) original approach used a full information maximum likelihood method with bivariate normal distribution assumptions for (ε_q, η_q) and (ε_q, ξ_q). Lee (1983) generalized Heckman's approach by allowing the univariate error terms ε_q, η_q, and ξ_q to be non-normal, using a technique to transform non-normal variables into normal variates, and then adopting a bivariate normal distribution to couple the transformed normal variables. Thus, while maintaining an efficient full-information likelihood approach, Lee's method relaxes the normality assumption on the marginals but still imposes a bivariate normal coupling. In addition to these full-information likelihood methods, there are also two-step and more robust parametric approaches that impose a specific form of linearity between the error term in the discrete choice and the continuous outcome

[1] The reader will note that it is not possible to identify any dependence parameters between (η_q, ξ_q) because the vehicle miles of travel is observed in only one of the two regimes for any given household.

(rather than a pre-specified bivariate joint distribution). These approaches are based on the Heckman method for the binary choice case, which was generalized by Hay (1980) and Dubin and McFadden (1984) for the multinomial case. The approach involves the first step estimation of the discrete choice equation given distributional assumptions on the choice model error terms, followed by the second step estimation of the continuous equation after the introduction of a correction term that is an estimate of the expected value of the continuous equation error term given the discrete choice. However, these two-step methods do not perform well when there is a high degree of collinearity between the explanatory variables in the choice equation and the continuous outcome equation, as is usually the case in empirical applications. This is because the correction term in the second step involves a non-linear function of the discrete choice explanatory variables. But this non-linear function is effectively a linear function for a substantial range, causing identification problems when the set of discrete choice explanatory variables and continuous outcome explanatory variables are about the same. The net result is that the two-step approach can lead to unreliable estimates for the outcome equation (see Leung and Yu, 2000 and Puhani, 2000).

Overall, Lee's full information maximum likelihood approach has seen more application in the literature relative to the other approaches just described because of its simple structure, ease of estimation using a maximum likelihood approach, and its lower vulnerability to the collinearity problem of two-step methods. But Lee's approach is also critically predicated on the bivariate normality assumption on the transformed normal variates in the discrete and continuous equation, which imposes the restriction that the dependence between the transformed discrete and continuous choice error terms is linear and symmetric. There are two ways that one can relax this joint bivariate normal coupling used in Lee's approach. One is to use semi-parametric or non-parametric approaches to characterize the relationship between the discrete and continuous error terms, and the second is to test alternative copula-based bivariate distributional assumptions to couple error terms. Each of these approaches is discussed in turn next.

1.1 Semi-Parametric and Non-Parametric Approaches

The potential econometric estimation problems associated with Lee's parametric distribution approach has spawned a whole set of semi-parametric and non-parametric two-step estimation methods to handle sample selection, apparently having beginnings in the semi-parametric work of Heckman and Robb (1985). The general approach in these methods is to first estimate the discrete choice model in a semi-parametric or non-parametric fashion using methods developed by, among others, Cosslett (1983), Ichimura (1993), Matzkin (1992, 1993), and Briesch *et al.* (2002). These estimates then form the basis to develop an index function to generate a correction term in the continuous equation that is an estimate of the expected value of the continuous equation error term given the discrete choice. While in the two-step parametric methods, the index function is defined based on the assumed marginal and joint distributional assumptions, or on an assumed marginal distribution for the discrete choice along with a specific linear form of relationship between the discrete and continuous equation error terms, in the semi- and non-parametric approaches, the index function is approximated by a flexible function of parameters such as the polynomial, Hermitian, or Fourier series expansion methods (see Vella, 1998 and Bourguignon *et al.*, 2007 for good reviews). But, of course, there are "no free lunches". The semi-parametric and non-parametric approaches involve a large number of parameters to estimate, are relatively very inefficient from an econometric estimation standpoint, typically do not allow the testing and inclusion of a rich set of explanatory variables with the usual range of sample sizes available in empirical contexts, and are difficult to implement. Further, the computation of the covariance matrix of parameters for inference is anything but simple in the semi- and non-parametric approaches. The net result is that the semi- and non-parametric approaches have been pretty much confined to the academic realm and have seen little use in actual empirical application.

1.2 The Copula Approach

The turn toward semi-parametric and non-parametric approaches to dealing with sample selection was ostensibly because of a sense that replacing Lee's parametric bivariate normal coupling with alternative bivariate couplings would lead to substantial computational burden. However, an approach referred to as the "Copula" approach has recently revived interest in maintaining a Lee-like sample selection framework, while generalizing Lee's framework to adopt and test a whole set of alternative bivariate couplings that can allow non-linear and asymmetric dependencies. A copula is essentially a multivariate functional form for the joint distribution of random variables derived purely from pre-specified parametric marginal distributions of each random variable. The reasons for the interest in the copula approach for sample selection models are several. First, the copula approach does not entail any more computational burden than Lee's approach. Second, the approach allows the analyst to stay within the familiar maximum likelihood framework for estimation and inference, and does not entail any kind of numerical integration or simulation machinery. Third, the approach allows the marginal distributions in the discrete and continuous equations to take on any parametric distribution, just as in Lee's method. Finally, under the copula approach, Lee's coupling method is but one of a suite of different types of couplings that can be tested.

In this paper, we apply the copula approach to examine built environment effects on vehicle miles of travel (VMT). The rest of this paper is structured as follows. The next section provides a theoretical overview of the copula approach, and presents several important copula structures.

Section 3 discusses the use of copulas in sample selection models. Section 4 provides an overview of the data sources and sample used for the empirical application. Section 5 presents and discusses the modeling results. The final section concludes the paper by highlighting paper findings and summarizing implications.

2. OVERVIEW OF THE COPULA APPROACH

2.1 Background

The incorporation of dependency effects in econometric models can be greatly facilitated by using a copula approach for modeling joint distributions, so that the resulting model can be in closed-form and can be estimated using direct maximum likelihood techniques (the reader is referred to Trivedi and Zimmer, 2007 or Nelsen, 2006 for extensive reviews of copula theory, approaches, and benefits). The word copula itself was coined by Sklar, 1959 and is derived from the Latin word "copulare", which means to tie, bond, or connect (see Schmidt, 2007). Thus, a copula is a device or function that generates a stochastic dependence relationship (*i.e.*, a multivariate distribution) among random variables with pre-specified marginal distributions. In essence, the copula approach separates the marginal distributions from the dependence structure, so that the dependence structure is entirely unaffected by the marginal distributions assumed. This provides substantial flexibility in correlating random variables, which may not even have the same marginal distributions.

The effectiveness of a copula approach has been recognized in the statistics field for several decades now (see Schweizer and Sklar, 1983, Ch. 6), but it is only recently that copula-based methods have been explicitly recognized and employed in the finance, actuarial science, hydrological modeling, and econometrics fields (see, for example, Embrechts *et al.*, 2002, Cherubini *et al.*, 2004, Frees and Wang, 2005, Genest and Favre, 2007, Grimaldi and Serinaldi, 2006, Smith, 2005, Prieger, 2002, Zimmer and Trivedi, 2006, Cameron *et al.*, 2004, Junker and May, 2005, and Quinn, 2007). The precise definition of a copula is that it is a multivariate distribution function defined over the unit cube linking uniformly distributed marginals. Let C be a K-dimensional copula of uniformly distributed random variables $U_1, U_2, U_3, \ldots, U_K$ with support contained in $[0,1]^K$. Then,

$$C_\theta\,(u_1, u_2, \ldots, u_K) = \Pr(U_1 < u_1, U_2 < u_2, \ldots, U_K < u_K), \tag{2}$$

where θ is a parameter vector of the copula commonly referred to as the dependence parameter vector. A copula, once developed, allows the generation of joint multivariate distribution functions with given marginals. Consider K random variables $Y_1, Y_2, Y_3, \ldots, Y_K$, each with univariate continuous marginal distribution functions $F_k(y_k) = \Pr(Y_k < y_k)$, k =1, 2, 3, …, K. Then, by the integral transform result, and using the notation $F_k^{-1}(.)$ for the inverse univariate cumulative distribution function, we can write the following expression for each k (k = 1, 2, 3, …, K):

$$F_k(y_k) = \Pr(Y_k < y_k) = \Pr(F_k^{-1}(U_k) < y_k) = \Pr(U_k < F_k(y_k)). \tag{3}$$

Then, by Sklar's (1973) theorem, a joint K-dimensional distribution function of the random variables with the continuous marginal distribution functions $F_k(y_k)$ can be generated as follows:

$$\begin{aligned} F(y_1, y_2, \ldots, y_K) &= \Pr(Y_1 < y_1, Y_2 < y_2, \ldots, Y_K < y_K) = \Pr(U_1 < F_1(y_1),, U_2 < F_2(y_2), \ldots, U_K < F_K(y_K)) \\ &= C_\theta\,(u_1 = F_1(y_1), u_2 = F_2(y_2), \ldots, u_K = F_K(y_K)). \end{aligned} \tag{4}$$

Conversely, by Sklar's theorem, for any multivariate distribution function with continuous marginal distribution functions, a unique copula can be defined that satisfies the condition in Equation (4).

Copulas themselves can be generated in several different ways, including the method of inversion, geometric methods, and algebraic methods (see Nelsen, 2006; Ch. 3). For instance, given a known multivariate distribution $F(y_1, y_2, \ldots, y_K)$ with continuous margins $F_k(y_k)$, the inversion method inverts the relationship in Equation (4) to obtain a copula:

$$\begin{aligned} C_\theta(u_1, u_2, \ldots, u_K) &= \Pr(U_1 < u_1, U_2 < u_2, \ldots, U_K < u_K) \\ &= \Pr(Y_1 < F^{-1}{}_1(u_1), Y_2 < F^{-1}{}_2(u_2), \ldots, Y_3 < F^{-1}{}_3(u_3)) \\ &= F(y_1 = F^{-1}{}_1(u_1), y_2 = F^{-1}{}_2(u_2), \ldots, y_K = F^{-1}{}_k(u_k)). \end{aligned} \qquad (5)$$

Once the copula is developed, one can revert to Equation (4) to develop new multivariate distributions with arbitrary univariate margins.

A rich set of copula types have been generated using the inversion and other methods, including the Gaussian copula, the Farlie-Gumbel-Morgenstern (FGM) copula, and the Archimedean class of copulas (including the Clayton, Gumbel, Frank, and Joe copulas). These copulas are discussed later in the context of bivariate distributions. In such bivariate distributions, while θ can be a vector of parameters, it is customary to use a scalar measure of dependence. In the next section, we discuss some copula properties and dependence structure concepts for bivariate copulas, though generalizations to higher dimensions are possible.

2.2 Copula Properties and Dependence Structure

Consider any bivariate copula $C_\theta(u_1, u_2)$. Since this is a bivariate cumulative distribution function, the copula should satisfy the well known Fréchet-Hoeffding bounds (see Kwerel, 1988). Specifically, the Fréchet lower bound $W(u_1, u_2)$ is $\max(u_1 + u_2 - 1, 0)$ and the Fréchet upper bound $M(u_1, u_2)$ is $\min(u_1, u_2)$. Thus,

$$W(u_1, u_2) \le C_\theta(u_1, u_2) \le M(u_1, u_2). \qquad (6)$$

From Sklar's theorem of Equation (4), we can also re-write the equation above in terms of Fréchet bounds for the multivariate distribution $F(y_1, y_2)$ generated from the copula $C_\theta(u_1, u_2)$:

$$\max(F_1(y_1) + F_2(y_2) - 1, 0) \le F(y_1, y_2) \le \min(F_1(y_1), F_2(y_2)). \qquad (7)$$

If the copula $C_\theta(u_1, u_2)$ is equal to the lower bound $W(u_1, u_2)$ in Equation (6), or equivalently if $F(y_1, y_2)$ is equal to the lower bound in Equation (7), then the random variables Y_1 and Y_2 are almost surely decreasing functions of each other and are called "countermonotonic". On the other hand, if the copula $C_\theta(u_1, u_2)$ is equal to the upper bound $M(u_1, u_2)$ in Equation (6), or equivalently if $F(y_1, y_2)$ is equal to the upper bound in Equation (7), then the random variables Y_1 and Y_2 are almost surely increasing functions of each other and are called "comonotonic". The case when $C_\theta(u_1, u_2) = \Pi = u_1 u_2$, or equivalently $F(y_1, y_2) = F_1(y_1) F_2(y_2)$, corresponds to stochastic independence between Y_1 and Y_2.

Different copulas provide different levels of ability to capture dependence between Y_1 and Y_2 based on the degree to which they cover the interval between the Fréchet-Hoeffding bounds. Comprehensive copulas are those that (1) attain or approach the lower bound W as θ approaches the lower bound of its permissible range, (2) attain or approach the upper bound M as θ approaches its upper bound, and (3) cover the entire domain between W and M (including the product copula case Π as a special or limiting case). Thus, comprehensive copulas parameterize the full range of dependence as opposed to non-comprehensive copulas that are only able to capture dependence in a limited manner. As we discuss later, the Gaussian and Frank copulas are comprehensive in their dependence structure, while the FGM, Clayton, Gumbel, and Joe copulas are not comprehensive.

To better understand the generated dependence structures between the random variables (Y_1, Y_2) based on different copulas, and examine the coverage offered by non-comprehensive copulas, it is useful to construct a scalar dependence measure between Y_1 and Y_2 that satisfies four properties as listed below (see Embrechts *et al.*, 2002):

(1) $\delta(Y_1, Y_2) = \delta(Y_2, Y_1)$

(2) $-1 \leq \delta(Y_1, Y_2) \leq 1$ (8)

(3) $\delta(Y_1, Y_2) = 1 \Leftrightarrow (Y_1, Y_2)$ comonotonic; $\delta(Y_1, Y_2) = -1 \Leftrightarrow (Y_1, Y_2)$ countermonotonic

(4) $\delta(Y_1, Y_2) = \delta(G_1(Y_1), G_2(Y_2))$, where G_1 and G_2 are two (possibly different) strictly increasing transformations.

The traditional dependence concept of correlation coefficient ρ (*i.e.*, the Pearson's product-moment correlation coefficient) is a measure of linear dependence between Y_1 and Y_2. It satisfies the first two of the properties discussed above. However, it satisfies the third property only for bivariate elliptical distributions (including the bivariate normal distribution) and adheres to the fourth property only for strictly increasing <u>linear</u> transformations (see Embrechts *et al.*, 2002 for specific examples where the Pearson's correlation coefficient fails the third and fourth properties). In addition, $\rho = 0$ does not necessarily imply independence. A simple example given by Embrechts *et al.*, 2002 is that $\rho(Y_1, Y_2) = 0$ if $Y_1 \sim N\ (0,1)$ and $Y_2 = Y_1^2$, even though Y_1 and Y_2 are clearly dependent. This is because Cov(Y_1, Y_2) = 0 implies zero correlation, but the stronger condition that Cov($G_1(Y_1)$, ($G_2(Y_2)$) = 0 for any functions G_1 and G_2 is needed for zero dependence. Other limitations of the Pearson's correlation coefficient include that it is not informative for asymmetric distributions (Boyer *et al.*, 1999), effectively goes to zero as one asymptotically heads into tail events just because the joint distribution gets flatter at the tails (Embrechts *et al.*, 2002), and the attainable correlation coefficient values within the [–1, 1] range depend upon the margins $F_1(.)$ and $F_2(.)$.

The limitations of the traditional correlation coefficient have led statisticians to the use of concordance measures to characterize dependence. Basically, two random variables are labeled as being concordant (discordant) if large values of one variable are associated with large (small) values of the other, and small values of one variable are associated with small (large) values of the other. This concordance concept has led to the use of two measures of dependence in the literature: the Kendall's τ and the Spearman's ρ_S.

Kendall's τ measure of dependence between two random variables (Y_1, Y_2) is defined as the probability of concordance minus the probability of discordance. Notationally,

$$\tau(Y_1,Y_2)=P\Big((Y_1-\tilde{Y}_1)(Y_2-\tilde{Y}_2)>0\Big)-P\Big((Y_1-\tilde{Y}_1)(Y_2-\tilde{Y}_2)<0\Big), \tag{9}$$

where $(\tilde{Y}_1,\tilde{Y}_2)$ is an independent copy of (Y_1,Y_2). The first expression on the right side is the probability of concordance of (Y_1,Y_2) and $(\tilde{Y}_1,\tilde{Y}_2)$, and the second expression on the right side is the probability of discordance of the same two vectors. It is straightforward to show that if $C_\theta(u_1,u_2)$ is the copula for the continuous random variables (Y_1,Y_2), *i.e.*, if $F(y_1,y_2)=C_\theta(u_1=F_1(y_1),\ u_2=F_2(y_2))$, then the expression above collapses to the following (see Nelsen, 2006, page 159 for a proof):

$$\tau(Y_1,Y_2)=4\iint_{[0,1]^2}C_\theta(u_1,u_2)dC_\theta(u_1,u_2)-1=4E[C_\theta(U_1,U_2)]-1, \tag{10}$$

where the second expression is the expected value of the function $C_\theta(U_1,U_2)$ of uniformly distributed random variables U_1 and U_2 with a joint distribution function *C*.

Spearman's ρ_S measure of dependence between two random variables (Y_1,Y_2) is defined as follows. Let $(\tilde{Y}_1,\tilde{Y}_2)$ and $(\breve{Y}_1,\breve{Y}_2)$ be independent copies of (Y_1,Y_2). That is, (Y_1,Y_2), $(\tilde{Y}_1,\tilde{Y}_2)$, and $(\breve{Y}_1,\breve{Y}_2)$ are all independent random vectors, each with a common joint distribution function *F*(.,.) and margins F_1 and F_2. Then, Spearman's ρ_S is three times the probability of concordance minus the probability of discordance for the two vectors (Y_1,Y_2) and $(\tilde{Y}_1,\breve{Y}_2)$:

$$\rho_S(Y_1,Y_2)=3\Big(P\big((Y_1-\tilde{Y}_1)(Y_2-\breve{Y}_2)>0\big)-P\big((Y_1-\tilde{Y}_1)(Y_2-\breve{Y}_2)\big)<0\Big) \tag{11}$$

In the above expression, note that the distribution function for (Y_1,Y_2) is F(.,.), while the distribution function of $(\tilde{Y}_1,\breve{Y}_2)$ is $F_1(.)F_2(.)$. because of the independence of $\tilde{Y}_1$ and $\breve{Y}_2$. The coefficient "3" is a normalization constant, since the expression in parenthesis is bounded in the region [–1/3, 1/3] (see Nelsen, 2006, pg 161). In terms of the copula $C_\theta(u_1,u_2)$ for the continuous random variables (Y_1,Y_2), ρ_S can be simplified to the expression below:

$$\rho_S(Y_1,Y_2)=12\iint_{[0,1]^2}u_1u_2dC_\theta(u_1,u_2)-3=12\iint_{[0,1]^2}C_\theta(u_1,u_2)du_1du_2-3=12E[(U_{1,}U_2)]-3 \tag{12}$$

where $U_1=F_1(Y_1)$ and $U_2=F_2(Y_2)$ are uniform random variables with joint distribution function $C_\theta(u_1,u_2)$. Since U_1 and U_2 have a mean of 0.5 and a variance of 1/12, the expression above can be re-written as:

$$\begin{aligned}\rho_S(Y_1,Y_2)=12E[(U_1U_2)]-3&=\frac{E(U_1U_2)-1/4}{1/12}=\frac{E(U_1U_2)-E(U_1)E(U_2)}{\sqrt{Var(U_1)}\sqrt{Var(U_2)}}\\ &=\rho(F_1(Y_1),F_2(Y_2))\end{aligned} \tag{13}$$

Thus, the Spearman ρ_S dependence measure for a pair of continuous variables (Y_1,Y_2) is equivalent to the familiar Pearson's correlation coefficient ρ for the grades of Y_1 and Y_2, where the grade of Y_1 is $F_1(Y_1)$ and the grade of Y_2 is $F_2(Y_2)$.

The Kendall's τ and the Spearman's ρ_S measures can be shown to satisfy all the four properties listed in Equation (8). In addition, both assume the value of zero under independence and are not dependent on the margins $F_1(.)$ and $F_2(.)$. Hence, these two concordance measures are used to characterize dependence structures in the copula literature, rather than the familiar Pearson's correlation coefficient.

2.3 Alternative Copulas

Several copulas have been formulated in the literature, and these copulas can be used to tie random variables together. In the bivariate case, given a particular bivariate copula, a bivariate distribution $F(y_1, y_2)$ can be generated for two random variables Y_1 (with margin F_1) and Y_2 (with margin F_2) using the general expression of Equation (4) as:

$$F(y_1, y_2) = C_\theta(u_1 = F_1(y_1), u_2 = F_2(y_2)) \qquad (14)$$

For given functional forms of the margins, the precise bivariate dependence profile between the variables Y_1 and Y_2 is a function of the copula $C_\theta(u_1, u_2)$ used, and the dependence parameter θ. But, regardless of the margins assumed, the overall nature of the dependence between Y_1 and Y_2 is determined by the copula. Note also that the Kendall's τ and the Spearman's ρ_S measures are functions only of the copula used and the dependence parameter in the copula, and not dependent on the functional forms of the margins. Thus, bounds on the τ and ρ_S measures for any copula will apply to all bivariate distributions derived from that copula. In the rest of this section, we focus on bivariate forms of the Gaussian copula, the Farlie-Gumbel-Morgenstern (FGM) copula, and the Archimedean class of copulas. To visualize the dependence structure for each copula, we follow Nelsen (2006) and Armstrong (2003), and first generate 1000 pairs of uniform random variates from the copula with a specified value of Kendall's τ (see http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/Supp_material.pdf for details of the procedure to generate uniform variates from each copula). Then, we transform these uniform random variates to normal random variates using the integral transform result ($Y_1 = \Phi^{-1}(U_1)$ and $Y_2 = \Phi^{-1}(U_2)$). For each copula, we plot two-way scatter diagrams of the realizations of the normally distributed random variables Y_1 and Y_2. In addition, Table 1 provides comprehensive details of each of the copulas.

2.3.1 The Gaussian copula

The Gaussian copula is the most familiar of all copulas, and forms the basis for Lee's (1983) sample selection mechanism. The copula belongs to the class of elliptical copulas, since the Gaussian copula is simply the copula of the elliptical bivariate normal distribution (the density contours of elliptical distributions are elliptical with constant eccentricity). The Gaussian copula takes the following form:

$$C_\theta(u_1, u_2) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \theta), \tag{15}$$

where $\Phi_2(.,.,\theta)$ is the bivariate cumulative distribution function with Pearson's correlation parameter $\theta(-1 \le \theta \le 1)$. The Gaussian copula is comprehensive in that it attains the Fréchet lower and upper bounds, and captures the full range of (negative or positive) dependence between two random variables. However, it also assumes the property of asymptotic independence. That is, regardless of the level of correlation assumed, extreme tail events appear to be independent in each margin just because the density function gets very thin at the tails (see Embrechts *et al.*, 2002). Further, the dependence structure is radially symmetric about the center point in the Gaussian copula. That is, for a given correlation, the level of dependence is equal in the upper and lower tails.[2]

The Kendall's τ and the Spearman's ρ_S measures for the Gaussian copula can be written in terms of the dependence (correlation) parameter θ as $\tau = (2/\pi)\sin^{-1}(\theta)$ and

[2] Mathematically, the dependence structure of a copula is labeled as "radially symmetric" if the following condition holds: $C_\theta(u_1, u_2) = u_1 + u_2 - 1 + C_\theta(1 - u_1, 1 - u_2)$, where the right side of the expression above is the survival copula (see Nelsen, 2006, page 37). Consider two random variables Y_1 and Y_2 whose marginal distributions are individually symmetric about points a and b, respectively. Then, the joint distribution F of Y_1 and Y_2 will be radially symmetric about points a and b if and only if the underlying copula from which F is derived is radially symmetric.

$\rho_S = (6/\pi)\sin^{-1}(\theta/2)$, where $z = \sin^{-1}(\theta) \Rightarrow \sin(z) = \theta$. Thus, τ and ρ_S take on values on [–1, 1]. The Spearman's ρ_S tracks the correlation parameter closely.

A visual scatter plot of realizations from the Gaussian copula-generated distribution for transformed normally distributed margins is shown in Figure (1a). A value of $\tau = 0.75$ is used in the figure. Note that, for the Gaussian copula, the image is essentially the scatter plot of points from a bivariate normal distribution with a correlation parameter $\theta = 0.9239$ (because we are using normal marginals). One can note the familiar elliptical shape with symmetric dependence. As one goes toward the extreme tails, there is more scatter, corresponding to asymptotic independence. The strongest dependence is in the middle of the distribution.

2.3.2 The Farlie-Gumbel-Morgenstern (FGM) copula

The FGM copula was first proposed by Morgenstern (1956), and also discussed by Gumbel (1960) and Farlie (1960). It has been well known for some time in Statistics (see Conway, 1979, Kotz *et al.*, 2000; Section 44.13). However, until Prieger (2002), it does not seem to have been used in Econometrics. In the bivariate case, the FGM copula takes the following form:

$$C_\theta(u_1, u_2) = u_1 u_2 [1 + \theta(1 - u_1)(1 - u_2)]. \tag{16}$$

For the copula above to be 2-increasing (that is, for any rectangle with vertices in the domain of [0,1] to have a positive volume based on the function), θ must be in [–1, 1]. The presence of the θ term allows the possibility of correlation between the uniform marginals u_1 and u_2. Thus, the FGM copula has a simple analytic form and allows for either negative or positive dependence. Like the Gaussian copula, it also imposes the assumptions of asymptotic independence and radial symmetry in dependence structure.

However, the FGM copula is not comprehensive in coverage, and can accommodate only relatively weak dependence between the marginals. The concordance-based dependence measures for the FGM copula can be shown to be $\tau = \frac{2}{9}\theta$ and $\rho_S = \frac{1}{3}\theta$, and thus these two measures are bounded on $\left[-\frac{2}{9}, \frac{2}{9}\right]$ and $\left[-\frac{1}{3}, \frac{1}{3}\right]$, respectively.

The FGM scatterplot for the normally distributed marginal case is shown in Figure (1b), where Kendall's τ is set to the maximum possible value of 2/9 (corresponding to $\theta = 1$). The weak dependence offered by the FGM copula is obvious from this figure.

2.3.3 The Archimedean class of copulas

The Archimedean class of copulas is popular in empirical applications (see Genest and MacKay, 1986 and Nelsen, 2006 for extensive reviews). This class of copulas includes a whole suite of closed-form copulas that cover a wide range of dependency structures, including comprehensive and non-comprehensive copulas, radial symmetry and asymmetry, and asymptotic tail independence and dependence. The class is very flexible, and easy to construct. Further, the asymmetric Archimedean copulas can be flipped to generate additional copulas (see Venter, 2001).

Archimedean copulas are constructed based on an underlying continuous convex decreasing generator function φ from [0, 1] to [0, ∞] with the following properties: $\varphi(1) = 0, \varphi'(t) < 0,$ and $\varphi''(t) > 0$ for all $0 < t < 1\, (\varphi'(t) = \partial\varphi / \partial t; \varphi''(t) = \partial^2\varphi / \partial^2 t).$ Further, in the discussion here, we will assume that $\varphi(0) = \infty$, so that an inverse φ^{-1} exists. With these preliminaries, we can generate bivariate Archimedean copulas as:

$$C_\theta(u_1, u_2) = \varphi^{-1}[\varphi(u_1) + \varphi(u_2)], \tag{17}$$

where the dependence parameter θ is embedded within the generator function. Note that the above expression can also be equivalently written as:

$$\varphi[C_\theta(u_1,u_2)] = [\varphi(u_1)+\varphi(u_2)]\,. \tag{18}$$

Using the differentiation chain rule on the equation above, we obtain the following important result for Archimedean copulas that will be relevant to the sample selection model discussed in the next section:

$$\frac{\partial C_\theta(u_1,u_2)}{\partial u_2} = \frac{\varphi'(u_2)}{\varphi'[C_\theta(u_1,u_2)]},\ \text{where}\ \varphi'(t) = \partial\varphi(t)/\partial t\,. \tag{19}$$

The density function of absolutely continuous Archimedean copulas of the type discussed later in this section may be written as:

$$c_\theta(u_1,u_2) = -\frac{\varphi''\big(C(u_1,u_2)\big)\varphi'(u_1)\,\varphi'(u_2)}{\big[\varphi'\big(C(u_1,u_2)\big)\big]^3}\ . \tag{20}$$

Another useful result for Archimedean copulas is that the expression for Kendall's τ in Equation (10) collapses to the following simple form (see Embrechts *et al.*, 2002 for a derivation):

$$\tau = 1+4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt\,. \tag{21}$$

In the rest of this section, we provide an overview of four different Archimedean copulas: the Clayton, Gumbel, Frank, and Joe copulas.

2.3.3.1 The Clayton copula

The Clayton copula has the generator function $\varphi(t) = (1/\theta)(t^{-\theta}-1)$, giving rise to the following copula function (see Huard *et al.*, 2006):

$$C_\theta(u_1,u_2) = (u_1^{-\theta}+u_2^{-\theta}-1)^{-1/\theta},\ 0<\theta<\infty\,. \tag{22}$$

The above copula, proposed by Clayton (1978), cannot account for negative dependence. It attains the Fréchet upper bound as $\theta \to \infty$, but cannot achieve the Fréchet lower bound. Using the Archimedean copula expression in Equation (21) for τ, it is easy to see that τ is related to θ by $\tau = \theta/(\theta + 2)$, so that 0 < τ < 1 for the Clayton copula. Independence corresponds to $\theta \to 0$.

The figure corresponding to the Clayton copula for $\tau = 0.75$ indicates asymmetric and positive dependence [see Figure (1c)]. The tight clustering of the points in the left tail, and the fanning out of the points toward the right tail, indicate that the copula is best suited for strong left tail dependence and weak right tail dependence. That is, it is best suited when the random variables are likely to experience low values together (such as loan defaults during a recession). Note that the Gaussian copula cannot replicate such asymmetric and strong tail dependence at one end.

2.3.3.2 The Gumbel copula

The Gumbel copula, first discussed by Gumbel (1960) and sometimes also referred to as the Gumbel-Hougaard copula, has a generator function given by $\varphi(t) = (-\ln t)^{\theta}$. The form of the copula is provided below:

$$C_{\theta}(u_1, u_2) = \exp\left(-\left[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}\right]^{1/\theta}\right), \ 1 \le \theta < \infty. \tag{23}$$

Like the Clayton copula, the Gumbel copula cannot account for negative dependence, but attains the Fréchet upper bound as $\theta \to \infty$. Kendall's τ is related to θ by $\tau = 1 - (1/\theta)$, so that 0 < τ < 1, with independence corresponding to $\theta = 1$.

As can be observed from Figure (1d), the Gumbel copula for $\tau = 0.75$ has a dependence structure that is the reverse of the Clayton copula. Specifically, it is well suited for the case when there is strong right tail dependence (strong correlation at high values) but weak left tail dependence (weak correlation at low values). However, the contrast between the dependence in the two tails of the Gumbel is clearly not as pronounced as in the Clayton.

2.3.3.3 The Frank copula

The Frank copula, proposed by Frank (1979), is the only Archimedean copula that is comprehensive in that it attains both the upper and lower Fréchet bounds, thus allowing for positive and negative dependence. It is radially symmetric in its dependence structure and imposes the assumption of asymptotic independence. The generator function is $\varphi(t) = -\ln[(e^{-\theta t} - 1)/(e^{-\theta} - 1)]$, and the corresponding copula function is given by:

$$C_\theta(u_1, u_2) = -\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right), \; -\infty < \theta < \infty . \tag{24}$$

Kendall's τ does not have a closed form expression for Frank's copula, but may be written as (see Nelsen, 2006, pg 171):

$$\tau = 1 - \frac{4}{\theta}\left[1 - D_F(\theta)\right], D_F(\theta) = \frac{1}{\theta}\int_{t=0}^{\theta} \frac{t}{e^t - 1} dt . \tag{25}$$

The range of τ is –1 < τ < 1. Independence is attained in Frank's copula as $\theta \to 0$.

The scatter plot for points from the Frank copula is provided in Figure (1e) for a value of $\tau = 0.75$, which translates to a θ value of 14.14. The points show very strong central dependence (even stronger than the Gaussian copula, as can be noted from the substantial central clustering) and very weak tail dependence (even weaker than the Gaussian copula, as can be noted from the fanning out at the tails). Thus, the Frank copula is suited for very strong central dependency with very weak tail dependency. The Frank copula has been used quite extensively in empirical applications (see Meester and MacKay, 1994; Micocci and Masala, 2003).

2.3.3.4 The Joe copula

The Joe copula, introduced by Joe (1993, 1997), has a generator function $\varphi(t) = -\ln[1-(1-t)^{\theta}]$ and takes the following copula form:

$$C_{\theta}(u_1,u_2) = 1-\left[(1-u_1)^{\theta}+(1-u_2)^{\theta}-(1-u_1)^{\theta}(1-u_2)^{\theta}\right]^{1/\theta},\ 1 \le \theta < \infty. \tag{26}$$

The Joe copula is similar to the Clayton copula. It cannot account for negative dependence. It attains the Fréchet upper bound as $\theta \to \infty$, but cannot achieve the Fréchet lower bound. The relationship between τ and θ for Joe's copula does not have a closed form expression, but takes the following form:

$$\tau = 1+\frac{4}{\theta}D_J(\theta), D_J(\theta) = \int_{t=0}^{1} \frac{\left[\ln(1-t^{\theta})\right](1-t^{\theta})}{t^{\theta-1}}dt. \tag{27}$$

The range of τ is between 0 and 1, and independence corresponds to $\theta = 1$.

Figure (1f) presents the scatter plot for the Joe copula (with $\tau = 0.75$), which indicates that the Joe copula is similar to the Gumbel, but the right tail positive dependence is stronger (as can be observed from the tighter clustering of points in the right tail). In fact, from this standpoint, the Joe copula is closer to being the reverse of the Clayton copula than is the Gumbel.

3. MODEL ESTIMATION AND MEASUREMENT OF TREATMENT EFFECTS

In the current paper, we introduce copula methods to accommodate residential self-selection in the context of assessing built environments effects on travel choices. To our knowledge, this is the first consideration and application of the copula approach in the urban planning and transportation literature (see Prieger, 2002 and Schmidt, 2003 for the application of copulas in the Economics literature). In the next section, we discuss the maximum likelihood estimation approach for estimating the parameters of Equation system (1) with different copulas.

3.1 Maximum Likelihood Estimation

Let the univariate standardized marginal cumulative distribution functions of the error terms $(\varepsilon_q, \eta_q, \xi_q)$ in Equation (1) be $(F_\varepsilon, F_\eta, F_\xi)$, respectively. Assume that η_q has a scale parameter of σ_η, and ξ_q has a scale parameter of σ_ξ. Also, let the standardized joint distribution of (ε_q, η_q) be *F*(.,.) with the corresponding copula $C_{\theta_0}(.,.)$, and let the standardized joint distribution of (ε_q, ξ_q) be *G*(.,.) with the corresponding copula $C_{\theta_1}(.,.)$.

Consider a random sample size of *Q* (*q*=1,2,…,*Q*) with observations on $(r_q, m_{q0}, m_{q1}, x_q, z_q, w_q)$. The switching regime model has the following likelihood function (see Appendix A for the derivation).

$$L = \prod_{q=1}^{Q} \left[\frac{1}{\sigma_\eta} \cdot f_\eta\left(\frac{m_{q0} - \alpha' z_q}{\sigma_\eta} \right) \cdot \frac{\partial}{\partial u_{q2}^0} C_{\theta_0}\left(u_{q1}^0, u_{q2}^0 \right) \right]^{1-r_q} \times \left[\frac{1}{\sigma_\xi} \cdot f_\xi\left(\frac{m_{q1} - \gamma' w_q}{\sigma_\xi} \right) \left\{ 1 - \frac{\partial}{\partial u_{q2}^1} C_{\theta_1}\left(u_{q1}^1, u_{q2}^1 \right) \right\} \right]^{r_q}, \qquad (28)$$

where $u_{q1}^0 = F_\varepsilon(-\beta' x_q),\ u_{q2}^0 = F_\eta\left(\frac{m_{q0} - \alpha' z_q}{\sigma_\eta} \right),\ u_{q1}^1 = u_{q1}^0,\ u_{q2}^1 = F_\xi\left(\frac{m_{q1} - \gamma' w_q}{\sigma_\xi} \right).$

Any copula function can be used to generate the bivariate dependence between (ε_q, η_q) and (ε_q, ξ_q), and the copulas can be different for these two dependencies (*i.e.*, C_{θ_0} and C_{θ_1} need not be the same). Thus, there is substantial flexibility in specifying the dependence structure, while still staying within the maximum likelihood framework and not needing any simulation machinery. In the current paper, we use normal distribution functions for the marginals $F_\varepsilon(.)$, $F_\eta(.)$ and $F_\xi(.)$, and test various different copulas for C_{θ_0} and C_{θ_1}. In Table 2, we provide the expression for $\frac{\partial}{\partial u_2} C_\theta(u_1, u_2)$

for the six copulas discussed in Section 2.3. For Archimedean copulas, the expression has the simple form provided in Equation (19).

The maximum-likelihood estimation of the sample selection model with different copulas leads to a case of non-nested models. The most widely used approach to select among the competing non-nested copula models is the Bayesian Information Criterion (or BIC; see Quinn, 2007, Genius and Strazzera, 2008, Trivedi and Zimmer, 2007, page 65). The BIC for a given copula model is equal to $-2\ln(L)+K\ln(Q)$, where $\ln(L)$ is the log-likelihood value at convergence, K is the number of parameters, and Q is the number of observations. The copula that results in the lowest BIC value is the preferred copula. But, if all the competing models have the same exogenous variables and a single copula dependence parameter θ, the BIC information selection procedure measure is equivalent to selection based on the largest value of the log-likelihood function at convergence.

3.2 Treatment Effects

The observed data for each household in the switching model of Equation (1) is its chosen residence location and the VMT given the chosen residential location. That is, we observe if $r_q = 0$ or $r_q = 1$ for each q, so that either m_{q0} or m_{q1} is observed for each q. We do not observe the data pair (m_{q0}, m_{q1}) for any household q. However, using the switching model, we would like to assess the impact of the neighborhood on VMT. In the social science terminology, we would like to evaluate the expected gains (*i.e.*, VMT increase) from the receipt of treatment (*i.e.*, residing in a conventional neighborhood). Heckman and Vytlacil, 2000 and Heckman *et al*., 2001 define a set of measures to study the influence of treatment, two important such measures being Average Treatment Effect (ATE) and the Effect of Treatment on the Treated (TT). We discuss these measures below, and propose two new measures labeled “Effect of Treatment on the Non-Treated (TNT)” and “Effect of Treatment on the Treated and Non-treated (TTNT)”. The mathematical expressions for an estimate of each measure are provided in Appendix B.

The ATE measure provides the expected VMT increase for a random household if it were to reside in a conventional neighborhood as opposed to a neo-urbanist neighborhood. The “Treatment on the Treated” or TT measure captures the expected VMT increase for a household randomly picked from the pool located in a conventional neighborhood if it were instead located in a neo-urbanist neighborhood (in social science parlance, it is the average impact of “treatment on the treated”; see Heckman and Vytlacil, 2005). In the current empirical setting, it is also of interest to assess the expected VMT increase for a household randomly picked from the pool located in a neo-urbanist neighborhood if it were instead located in a conventional neighborhood (*i.e.*, the “average impact of treatment on the non-treated” or TNT). Finally, one can combine the TT and TNT measures into a single measure that represents the average impact of treatment on the (currently) treated and (currently) non-treated (TTNT). In the current empirical context, it is the expected VMT change for a randomly picked household if it were relocated from its current neighborhood type to the other neighborhood type, measured in the *common direction* of change from a traditional neighborhood to a conventional neighborhood. The TTNT measure, in effect, provides the average expected change in VMT if all households were located in a conventional neighborhood relative to if all households were located in a neo-urbanist neighborhood. It includes both the “true” causal effect of neighborhood effects on VMT as well as the “self-selection” effect of households choosing neighborhoods based on their travel desires. The closer *TTNT* is to ATE, the lesser is the self-selection effect. Of course, in the limit that there is no self-selection, TTNT collapses to the ATE.

4. THE DATA

4.1 Data Sources

The data used for this analysis is drawn from the 2000 San Francisco Bay Area Household Travel Survey (BATS) designed and administered by MORPACE International Inc. for the Bay Area Metropolitan Transportation Commission (MTC). In addition to the 2000 BATS data, several other secondary data sources were used to derive spatial variables characterizing the activity-travel and built environment in the region. These included: (1) Zonal-level land-use/demographic coverage data, obtained from the MTC, (2) GIS layers of sports and fitness centers, parks and gardens, restaurants, recreational businesses, and shopping locations, obtained from the InfoUSA business directory, (3) GIS layers of bicycling facilities, obtained from MTC, and (4) GIS layers of the highway network (interstate, national, state and county highways) and the local roadways network (local, neighborhood, and rural roads), extracted from the Census 2000 Tiger files. From these secondary data sources, a wide variety of built environment variables were developed for the purpose of classifying the residential neighborhoods into neo-urbanist and conventional neighborhoods.

4.2 The Dependent Variables

This study uses factor analysis and a clustering technique to define a binary residential location variable that classifies the Traffic Analysis Zones (TAZs) of the Bay Area into neo-urbanist and conventional neighborhoods based on built environment measures. Factor analysis helps in reducing the correlated attributes (or *factors*) that characterize the built environment of a neighborhood into a manageable number of *principal components* (or variables). The clustering technique employs these *principal components* to classify zones into neo-urbanist or conventional neighborhoods. In the current paper, we employ the results from Pinjari *et al*. (2008) that identified two principal components to characterize the built environment of a zone - (1) Residential density and transportation/land-use environment, and (2) Accessibility to activity centers. The factors loading on the first component included bicycle lane density, number of zones accessible from the home zone by bicycle, street block density, household population density, and fraction of residential land use in the zone. The factors loading on the second component included bicycle lane density and number of physically active and natural recreation centers in the zone. The two principal components formed the basis for a cluster analysis that categorizes the 1099 zones in the Bay area into neo-urbanist or conventional neighborhoods (see Pinjari *et al*., 2008 for complete details). This binary variable is used as the dependent variable in the selection equation of Equation (1).

The continuous outcome dependent variable in each of the neo-urbanist and conventional neighborhood residential location regimes is the household vehicle miles of travel (VMT). This was obtained from the reported odometer readings before and after the two days of the survey for each vehicle in the household. The two-day vehicle-specific VMT was aggregated across all vehicles in the household to obtain a total two-day household VMT, which was subsequently averaged across the two survey days to obtain an average daily household VMT. The logarithm of the average daily household VMT was then used as the dependent variable, after recoding the small share (<5%) of households with a VMT value of zero to one (so that the logarithm of VMT takes a value of zero for these households).

The final estimation sample in our analysis includes 3696 households from 5 counties (San Francisco, San Mateo, Santa Clara, Alameda, and Contra Costa) of the Bay area. Among these households, about 34% of the households reside in neo-urbanist neighborhoods and 66% reside in conventional neighborhoods. The average daily household VMT is about 37 miles for households in neo-urbanist neighborhoods, and 68 miles for households in conventional neighborhoods.

5. EMPIRICAL ANALYSIS

5.1 Variables Considered

Several categories of variables were considered in the analysis, including household demographics, employment characteristics, and neighborhood characteristics. The neighborhood characteristics considered include population density, employment density, Hansen-type accessibility measures (such as accessibility to employment and accessibility to shopping; see Bhat and Guo, 2007 for the precise functional form), population by ethnicity in the neighborhood, presence/number of schools and physically active centers, and density of bicycle lanes and street blocks. These measures are included in the VMT outcome equation and capture the effect of variations in built environment across zones within each group of neo-urbanist and conventional neighborhoods.

5.2 Estimation Results

The empirical analysis involved estimating models with the same structure for (ε_q, η_q) and (ε_q, ξ_q), as well as different copula-based dependency structures. This led to 6 models with the same copula dependency structure (corresponding to the six copulas discussed in Section 2.3), and 24 models with different combinations of the six copula dependency structures for (ε_q, η_q) and (ε_q, ξ_q). We also estimated a model that assumed independence between ε_q and η_q, and ε_q and ξ_q.

The Bayesian Information Criterion, which collapses to a comparison of the log-likelihood values across different models, is employed to determine the best copula dependency structure combination. The log-likelihood values for the five best copula dependency structure combinations are: (1) Frank-Frank (-6842.2), (2) Frank-Joe (-6844.2), (3) FGM-Joe (-6851.0), (4) Independent-Joe (-6863.7), and (5) FGM-Gumbel (-6866.2). It is evident that the log-likelihood at convergence of the Frank-Frank and Frank-Joe copula combinations are higher compared to the other copula combinations. Between the Frank-Frank and Frank-Joe copula combinations, the former is slightly better. The log-likelihood value for the structure that assumes independence (*i.e.,* no self-selection effects) is -6878.1. All the five copula-based dependency models reject the independence assumption at any reasonable level of significance, based on likelihood ratio tests, indicating the significant presence of self-selection effects. Interestingly, however, the log-likelihood value at convergence for the classic textbook structure that assumes a Gaussian-Gaussian copula combination is -6877.9, indicating that there is no statistically significant difference between the Gaussian-Gaussian (G-G) and the independence-independence (I-I) copula structures. This is also observed in the estimated bivariate normal correlation parameters, which are -0.020 (t-statistic of 0.18) for the residential choice-neo-urbanist VMT regime error correlation and -0.050 (t-statistic of -0.50) for the residential choice-conventional neighborhood VMT regime error correlation. Clearly, the traditional G-G copula combination indicates the absence of self-selection effects. However, this is simply an artifact of the normal dependency structure, and is indicative of the kind of incorrect results that can be obtained by placing restrictive distributional assumptions.

In the following presentation of the empirical results, we focus our attention on the results of the Independent-Independent (or I-I copula) specification that ignores self-selection effects entirely and the Frank-Frank (or F-F copula) specification that provides the best data fit. Table 3 provides the results, which are discussed below.

5.2.1 Binary choice component

The results of the binary discrete equation of neighborhood choice provide the effects of variables on the propensity to reside in a conventional neighborhood relative to a neo-urbanist neighborhood. The parameter estimates indicate that younger households (*i.e.*, households whose heads are less than 35 years of age) are less likely to reside in conventional neighborhoods and more likely to reside in neo-urbanist neighborhoods, perhaps because of higher environmental sensitivity and/or higher need to be close to social and recreational activity opportunities (see also Lu and Pas, 1999). Households with children have a preference for conventional neighborhoods, potentially because of a perceived better quality of life/schooling for children in conventional neighborhoods compared to neo-urbanist neighborhoods. Also, as expected, households who own their home and who live in a single family dwelling unit are more likely to reside in conventional neighborhoods.

5.2.2 Log(VMT) continuous component for neo-urbanist neighborhood regime

The estimation results corresponding to the natural logarithm of vehicle miles of travel (VMT) in a neo-urbanist neighborhood highlight the significance of the number of household vehicles and number of full-time students. As expected, both of these effects are positive. In particular, log(VMT) increases with number of vehicles in the household and number of students. The effect of number of vehicles is non-linear, with a jump in log(VMT) for an increase from no vehicles to one vehicle, and a lesser impact for an increase from one vehicle to 2 or more vehicles (there were only two households in neo-urbanist neighborhoods with 3 vehicles, so we are unable to estimate impacts of vehicle increases beyond 2 vehicles in neo-urbanist neighborhoods). Interestingly, we did not find any statistically significant effect of employment and neighborhood characteristics, in part because the variability of these characteristics across households in neo-urbanist zones is relatively small.

The copula dependency parameter between the discrete choice residence error term and the log(VMT) error term for neo-urbanist households is highly statistically significant and negative for the F-F model. The θ estimate translates to a Kendall's τ value of -0.26. The negative dependency parameter indicates that a household that has a higher inclination to locate in conventional neighborhoods would travel less than an observationally equivalent "random" household if both these households were located in a neo-urbanist neighborhood (a "random" household, as used above, is one that is indifferent between residing in a neo-urbanist or a conventional neighborhood, based on factors unobserved to the analyst). Equivalently, the implication is that a household that makes the choice to reside in a neo-urbanist neighborhood is likely to travel more than an observationally equivalent random household in a neo-urbanist environment, and much more than if an observationally equivalent household from a conventional neighborhood were relocated to a neo-urbanist neighborhood. This may be attributed to, among other things, such unobserved factors characterizing households inclined to reside in neo-urbanist settings as a higher degree of comfort level driving in dense, one-way street-oriented, parking-loaded, traffic conditions.

The lower travel tendency of a random household in a neo-urbanist neighborhood (relative to a household that expressly chooses to locate in a neo-urbanist neighborhood) is teased out and reflected in the high statistically significant negative constant in the F-F copula model. On the other hand, the I-I model assumes, incorrectly, that the travel of households choosing to reside in neo-urbanist neighborhoods is independent of the choice of residence. The result is an inflation of the VMT generated by a random household if located in a neo-urbanist setting.

5.2.3 Log(VMT) continuous component for conventional neighborhood regime

The household socio-demographics that influence vehicle mileage for households in a conventional neighborhood include number of household vehicles, number of full-time students, and number of employed individuals. As expected, the effects of all of these variables are positive. The household vehicle effect is non-linear, with the marginal increase in log(VMT) decreasing with the number of vehicles. In addition, two neighborhood characteristics – density of vehicle lanes and accessibility to shopping – have statistically significant effects on log(VMT) in the conventional neighborhood regime. Both these effects are negative, as expected.

The dependency parameter in this segment for the F-F model is highly statistically significant and positive. The θ estimate translates to a Kendall's τ value of 0.36. The positive dependency indicates that a household that has a higher inclination to locate in conventional neighborhoods is likely to travel more in that setting than an observationally equivalent random household. Again, the I-I model ignores this residential self-selection in the estimation sample, resulting in an over-estimation of the VMT generated by a random household if located in a conventional neighborhood setting (see the higher constant in the I-I model relative to the F-F model corresponding to the conventional neighborhood VMT regime).

5.3 Treatment Effects

It is clear from the previous section that there are statistically significant residential self-selection effects; that is, households' choice of residence is linked to their VMT. To understand the magnitude of self-selection effects, we present point estimates of the treatment effects in this section. In addition to the point treatment effects (see Appendix B for the formulas), we also estimate large sample standard errors for the treatment effects using 1000 bootstrap draws. This involves drawing from the asymptotic distributions of parameters appearing in the treatment effect, and computing the standard deviation of the simulated treatment effect values.

The results are presented in Table 4 for the Independence-Independence (I-I) model and the three copula models with the best data fit, corresponding to the FGM-Joe (FG-J), Frank-Joe (F-J), and the Frank -Frank (F-F) copula models. Of course, the results from the traditional Gaussian-Gaussian (G-G) model are literally identical to the results from the I-I model, since the correlation parameters in the G-G model are small and very insignificant. The results show substantial variation in the treatment measures across models, except for the F-J and F-F models which provide similar results (this is not surprising, since the model parameters and log-likelihood values at convergence for these two models are almost the same, as discussed earlier in Section 5.2). According to the I-I model, a randomly selected household will have about the same VMT regardless of whether it is located in a conventional or neo-urbanist neighborhood (see the small and statistically insignificant ATE estimate for the I-I model). On the other hand, the other copula models indicate that there is indeed a statistically significant impact of the built environment on VMT. For instance, the best-fitting F-F model indicates that a randomly picked household will drive about 21 vehicle-miles per day more if in a conventional neighborhood relative to a neo-urbanist neighborhood. The important message here is that ignoring sample selection can lead to an underestimation or an overestimation of built environment effects (the general impression is that ignoring sample selection can only lead to an overestimation of built environment effects). Further, one needs to empirically test alternative copulas to determine which structure provides the best data fit, rather than testing the presence or absence of sample selection using normal dependency structures.

The results also show statistically significant variations in the other treatment effects between the I-I model and the non I-I models. The $\widehat{\text{TT}}$ and $\widehat{\text{TNT}}$ measures from the non I-I models reflect, as expected, that a household choosing to locate in a certain kind of neighborhood travels more in its chosen environment relative to an observationally equivalent random household. Thus, if a randomly picked household in a conventional neighborhood were to be relocated to a neo-urbanist neighborhood, the household's VMT is estimated to decrease by about 42 miles. Similarly, if a randomly picked household in a neo-urbanist neighborhood were to be relocated to a conventional neighborhood, the household's VMT is estimated to decrease by about 31 miles. On the other hand, if a randomly picked household that is indifferent to neighborhood type is moved from a conventional to a neo-urbanist neighborhood, the household's VMT is estimated to decrease by about 21 miles (which is, of course, the ATE measure).

The $\widehat{\text{TTNT}}$ measure is a weighted average of the $\widehat{\text{TT}}$ and $\widehat{\text{TNT}}$ measures, and shows that there would be a decrease of about 25 vehicle miles of travel per day if all households in the population (as represented by the estimation sample) were located in a neo-urbanist neighborhood rather than a conventional neighborhood. When compared to the average VMT of 58 miles, the implication is that one may expect a VMT reduction of about 43% by redesigning all neighborhoods to be of the neo-urbanist neighborhood type.[3] Finally, the $\widehat{\text{TTNT}}$ measure for the best F-F copula model shows that about 87% of the VMT difference between households residing in conventional and neo-urbanist neighborhoods is due to "true" built environment effects, while the remainder is due to residential self-selection effects. However, most importantly, it is critical to note that failure to accommodate the self-selection effect leads to a substantial underestimation of the "true" built

[3] Note that we are simply presenting this figure as a way to provide a magnitude effect of VMT reduction by designing urban environments to be of the neo-urbanist kind. In practice, different neighborhoods may be redesigned to different extents to make them less auto-dependent. Further, in a democratic society, demand will (and should) fuel supply. Thus, as long as there are individuals who prefer to live in a conventional setting, developers will provide that option.

environment effect (see the ATE for the I-I model of 0.49 miles relative to the ATE for the F-F model of 21.37 miles.

6. CONCLUSIONS AND IMPLICATIONS

In the current study, we apply a copula based approach to model residential neighborhood choice and daily household vehicle miles of travel (VMT) using the 2000 San Francisco Bay Area Household Travel Survey (BATS). The self-selection hypothesis in the current empirical context is that households select their residence locations based on their travel needs, which implies that observed VMT differences between households residing in neo-urbanist and conventional neighborhoods cannot be attributed entirely to built environment variations between the two neighborhoods types. A variety of copula-based models are estimated, including the traditional Gaussian-Gaussian (G-G) copula model. The results indicate that using a bivariate normal dependency structure suggests the absence of residential self-selection effects. However, other copula structures reveal a high and statistically significant level of residential self-selection, highlighting the potentially inappropriate empirical inferences from using incorrect dependency structures. In the current empirical case, we find the Frank-Frank (F-F) copula dependency structure to be the best in terms of data fit based on the Bayesian Information Criterion.

The examination of treatment effects provides very different implications from the traditional G-G copula model and the best F-F copula model. The first model effectively indicates that there are no self-selection effects and little to no effects of built environment on vehicle miles of travel. The F-F copula model indicates that the differences between VMT among neo-urbanist and conventional households are both due to self-selection as well as due to "true" built environment effects. Specifically, self-selection effects are estimated to constitute about 17% of the VMT difference between neo-urbanist and conventional households, while “true” built environment effects constitute the remaining 83% of the VMT difference.

In summary, this paper indicates the power of the copula approach to examine built environment effects on travel behavior, and to contribute to the debate on whether the empirically observed association between the built environment and travel behavior-related variables is a true reflection of underlying causality, or simply a spurious correlation attributable to the intervening relationship between the built environment and the characteristics of people who choose to live in particular built environments (or some combination of both these effects). The results of this study indicate that, in the empirical context of the current study, failure to accommodate residential self-selection effects can lead to a substantial mis-estimation of the true built environment effects. As importantly, the study indicates that use of a traditional normal bivariate distribution to characterize the relationship in errors between residential choice and VMT can lead to very misleading implications about built environment effects.

The copula approach used here can be extended to the case of sample selection with a multinomial treatment effect (see Spizzu *et al*., 2009 for a recent application). It should also have wide applicability in other bivariate/multivariate contexts in the transportation and other fields, including spatial dependence modeling (see Bhat and Sener, 2009).

ACKNOWLEDGMENTS

This research was funded in part by Environmental Protection Agency Grant R831837. The authors are grateful to Lisa Macias for her help in formatting this document. Two anonymous referees provided valuable comments on an earlier version of this paper.

REFERENCES

Armstrong, M., 2003. Copula catalogue - part 1: Bivariate archimedean copulas. Unpublished paper, Cerna, available at http://www.cerna.ensmp.fr/Documents/MA-CopulaCatalogue.pdf

Bhat, C.R., 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., Sener, I.N., 2009. A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

Bourguignon, S., Carfantan, H., Idier, J., 2007. A sparsity-based method for the estimation of spectral lines from irregularly sampled data. *IEEE Journal of Selected Topics in Signal Processing* 1(4), 575-585.

Boyer, B., Gibson, M., Loretan, M., 1999. Pitfalls in tests for changes in correlation. International Finance Discussion Paper 597, Board of Governors of the Federal Reserve System.

Briesch, R. A., Chintagunta, P. K., Matzkin, R. L., 2002. Semiparametric estimation of brand choice behavior. *Journal of the American Statistical Association* 97(460), 973-982.

Cameron, A. C., Li, T., Trivedi, P., Zimmer, D., 2004. Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal* 7(2), 566-584.

Cherubini, U., Luciano, E., Vecchiato, W., 2004. *Copula Methods in Finance*. John Wiley & Sons, Hoboken, NJ.

Clayton, D. G., 1978. A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence. *Biometrika* 65(1), 141-151.

Conway, D. A., 1979. Multivariate distributions with specified marginals. Technical Report #145, Department of Statistics, Stanford University.

Cosslett, S. R., 1983. Distribution-free maximum likelihood estimation of the binary choice model. *Econometrica* 51(3), 765-782.

Dubin, J. A., McFadden, D. L, 1984. An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52(1), 345-362.

Embrechts, P., McNeil, A. J., Straumann, D., 2002. Correlation and dependence in risk management: Properties and pitfalls. In M. Dempster (ed.) *Risk Management: Value at Risk and Beyond*, Cambridge University Press, Cambridge, 176-223.

Farlie, D. J. G., 1960. The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* 47(3-4), 307-323.

Frank, M. J., 1979. On the simultaneous associativity of F(x, y) and x + y - F(x, y). *Aequationes Mathematicae* 19(1), 194-226.

Frees, E. W., Wang, P. 2005. Credibility using copulas. *North American Actuarial Journal* 9(2), 31-48.

Genest, C., Favre, A.-C., 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* 12(4), 347-368.

Genest, C., MacKay, R. J., 1986. Copules archimediennes et familles de lois bidimensionnelles dont les marges sont donnees. *The Canadian Journal of Statistics* 14(2), 145-159.

Genius, M., Strazzera, E., 2008. Applying the copula approach to sample selection modeling. *Applied Economics* 40(11), 1443-1455.

Greene, W., 1981. Sample selection bias as a specification error: A comment. *Econometrica* 49(3), 795-798.

Grimaldi, S., Serinaldi, F., 2006. Asymmetric copula in multivariate flood frequency analysis. *Advances in Water Resources* 29(8), 1155-1167.

Gumbel, E. J., 1960. Bivariate exponential distributions. *Journal of the American Statistical Association* 55(292), 698-707.

Hay, J. W., 1980. Occupational choice and occupational earnings: Selectivity bias in a simultaneous logit-OLS model. Ph.D. Dissertation, Department of Economics, Yale University.

Heckman, J. (1974) Shadow prices, market wages and labor supply. *Econometrica*, 42(4), 679-694.

Heckman, J. (1976) The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5(4), 475-492.

Heckman, J. J., (1979) Sample selection bias as a specification error, *Econometrica*, 47(1), 153-161.

Heckman, J. J., 2001. Microdata, heterogeneity and the evaluation of public policy. *Journal of Political Economy* 109(4), 673-748.

Heckman, J. J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. In J. J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data,* Cambridge University Press, New York, 156-245.

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

Heckman, J. J., Vytlacil, E. J., 2005. Structural equations, treatment effects and econometric policy evaluation. *Econometrica* 73(3), 669-738.

Heckman, J. J., Tobias, J. L., Vytlacil, E. J., 2001. Four parameters of interest in the evaluation of social programs. *Southern Economic Journal* 68(2), 210-223.

Huard, D., Evin, G., Favre, A.-C., 2006. Bayesian copula selection. *Computational Statistics & Data Analysis* 51(2), 809-822.

Ichimura, H., 1993. Semiparametric Least Squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58(1-2), 71-120.

Joe, H., 1993. Parametric families of multivariate distributions with given marginals. *Journal of Multivariate Analysis* 46(2), 262-282.

Joe, H., 1997. *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

Junker, M., May, A., 2005. Measurement of aggregate risk with copulas. *The Econometrics Journal* 8(3), 428-454.

Kotz, S., Balakrishnan, N., Johnson, N. L., 2000. *Continuous Multivariate Distributions*, *Vol. 1, Models and Applications,* 2nd edition. John Wiley & Sons, New York.

Kwerel, S. M., 1988. Frechet bounds. In S. Kotz, N. L. Johnson (eds.) *Encyclopedia of Statistical Sciences*, Wiley & Sons, New York, 202-209.

Lee, L.-F., 1978. Unionism and wage rates: A simultaneous equation model with qualitative and limited dependent variables. *International Economic Review* 19(2), 415-433.

Lee, L.-F., 1982. Some approaches to the correction of selectivity bias. *Review of Economic Studies* 49(3), 355-372.

Lee, L.-F., 1983. Generalized econometric models with selectivity. *Econometrica* 51(2), 507-512.

Leung, S. F., Yu, S., 2000. Collinearity and two-step estimation of sample selection models: Problems, origins, and remedies. *Computational Economics* 15(3), 173-199.

Lu, X. L., Pas, E. I., 1999. Socio-demographics, activity participation, and travel behavior. *Transportation Research Part A* 33(1), 1-18.

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

Matzkin, R. L., 1992. Nonparametric and distribution-free estimation of the binary choice and the threshold crossing models. *Econometrica* 60(2), 239-270.

Matzkin, R. L., 1993. Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics* 58(1-2), 137-168.

Meester, S. G., MacKay, J., 1994. A parametric model for cluster correlated categorical data. *Biometrics* 50(4), 954-963.

Micocci, M., Masala, G., 2003. Pricing pension funds guarantees using a copula approach. Presented at AFIR Colloquium, International Actuarial Association, Maastricht, Netherlands.

Morgenstern, D., 1956. Einfache beispiele zweidimensionaler verteilungen. *Mitteilingsblatt fur Mathematische Statistik* 8(3), 234-235.

Nelsen, R. B., 2006. *An Introduction to Copulas* (2nd ed). Springer-Verlag, New York.

Pinjari, A. R., Eluru, N., Bhat, C. R., Pendyala, R. M., Spissu, E., 2008. Joint model of choice of residential neighborhood and bicycle ownership: Accounting for self-selection and unobserved heterogeneity. *Transportation Research Record* 2082, 17-26.

Prieger, J. E., 2002. A flexible parametric selection model for non-normal data with application to health care usage. *Journal of Applied Econometrics* 17(4), 367-392.

Puhani, P. A., 2000. The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* 14(1), 53-67.

Quinn, C., 2007. The health-economic applications of copulas: Methods in applied econometric research. Health, Econometrics and Data Group (HEDG) Working Paper 07/22, Department of Economics, University of York

Roy, A. D., 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, New Series 3(2), 135-146.

Schmidt, R., 2003. Credit risk modeling and estimation via elliptical copulae. In G. Bol, G. Nakhaeizadeh, S. T. Rachev, T. Ridder, and K.-H. Vollmer (eds.) *Credit Risk: Measurement, Evaluation, and Management*, 267-289, Physica-Verlag, Heidelberg.

Schmidt, T., 2007. Coping with copulas. In J. Rank (ed.) *Copulas - From Theory to Application in Finance*, 3-34, Risk Books, London.

Schweizer, B., Sklar, A., 1983. *Probabilistic Metric Spaces*. North-Holland, New York.

Sklar, A., 1959. Fonctions de répartition à *n* dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*, 8, 229-231.

Sklar, A., 1973. Random variables, joint distribution functions, and copulas. *Kybernetika* 9, 449-460.

Smith, M. D., 2005. Using copulas to model switching regimes with an application to child labour. *Economic Record* 81(S1), S47-S57.

Spissu, E., Pinjari, A. R., Pendyala, R. M., Bhat, C. R., 2009. A copula-based joint multinomial discrete-continuous model of vehicle type choice and miles of travel. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

Trivedi, P. K., Zimmer, D. M., 2007. Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics* 1(1), Now Publishers.

Vella, F., 1998. Estimating models with sample selection bias: A survey. *Journal of Human Resources* 33(1), 127-169.

Venter, G. G., 2001. Tails of copulas. Presented at ASTIN Colloquium, International Actuarial Association, Washington D.C.

Zimmer, D. M., Trivedi, P. K., 2006. Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *Journal of Business and Economic Statistics* 24(1), 63-76.

APPENDIX A

Using the notation in Section 3.1, the likelihood function may be written as:

$$L = \prod_{q=1}^{Q} \left[\left\{ \Pr[m_{q0} \mid r_q^* \le 0] \times \Pr[r_q^* \le 0] \right\}^{1-r_q} \times \left\{ \Pr[m_{q1} \mid r_q^* > 0] \times \Pr[r_q^* > 0] \right\}^{r_q} \right] \tag{A.1}$$

The conditional distributions in the expression above can be simplified. Specifically, we have the following:

$$\begin{aligned} \Pr[m_{q0} \mid r_q^* \le 0] &= \left\{ \Pr[r_q^* \le 0] \right\}^{-1} \times \frac{\partial}{\partial m_{q0}} F\left(-\beta' x_q , \frac{m_{q0} - \alpha' z_q}{\sigma_\eta} \right) \\ &= \left\{ \Pr[r_q^* \le 0] \right\}^{-1} \times \frac{1}{\sigma_\eta} \times \frac{\partial}{\partial t} F\left(-\beta' x_q , t \right) \Bigg|_{t = \frac{m_{q0} - \alpha' z_q}{\sigma_\eta}} \\ &= \left\{ \Pr[r_q^* \le 0] \right\}^{-1} \times \frac{1}{\sigma_\eta} \times \frac{\partial C_{\theta_0}(u_{q1}^0 , u_{q2}^0)}{\partial u_{q2}^0} \times f_\eta\left(\frac{m_{q0} - \alpha' z_q}{\sigma_\eta} \right) \end{aligned} \tag{A.2}$$

where $C_{\theta_0}(.,.)$ is the copula corresponding to *F* with $u_{q1}^0 = F_\varepsilon(-\beta' x_q)$ and $u_{q2}^0 = F_\eta\left(\frac{m_{q0} - \alpha' z_q}{\sigma_\eta} \right)$.

Similarly, we can write:

$$\begin{aligned} \Pr[m_{q1} \mid r_q^* > 0] &= \left\{ \Pr[r_q^* > 0] \right\}^{-1} \times \frac{\partial}{\partial m_{q1}} \left[F_\xi\left(\frac{m_{q1} - \gamma' w_q}{\sigma_\xi} \right) - G\left(-\beta' x_q , \frac{m_{q1} - \gamma' w_q}{\sigma_\xi} \right) \right] \\ &= \left\{ \Pr[r_q^* > 0] \right\}^{-1} \times \frac{1}{\sigma_\xi} \times \left[f_\xi\left(\frac{m_{q1} - \gamma' w_q}{\sigma_\xi} \right) - \frac{\partial}{\partial v} G\left(-\beta' x_q , v \right) \Bigg|_{v = \frac{m_{q1} - \gamma' w_q}{\sigma_\xi}} \right] \\ &= \left\{ \Pr[r_q^* > 0] \right\}^{-1} \times \frac{1}{\sigma_\xi} \left[f_\xi\left(\frac{m_{q1} - \gamma' w_q}{\sigma_\xi} \right) - \frac{\partial}{\partial u_{q2}^1} C_{\theta_1}(u_{q1}^1 , u_{q2}^1) \times f_\xi\left(\frac{m_{q1} - \gamma' w_q}{\sigma_\xi} \right) \right], \end{aligned} \tag{A3}$$

where $C_{\theta_1}(.,.)$ is the copula corresponding to *G* with $u_{q1}^1 = F_\varepsilon(-\beta' x_q)$ and $u_{q2}^1 = F_\xi\left(\frac{m_{q1} - \gamma' w_q}{\sigma_\xi} \right)$.

Substituting these conditional probabilities back into Equation (A.1) provides the general likelihood function expression for any sample selection model presented in Equation (28) in the text.

APPENDIX B. EXPRESSIONS FOR TREATMENT EFFECTS

$$\overset{\wedge}{\text{ATE}} = \frac{1}{Q}\sum_{q=1}^{Q}\left(\exp(\hat{\gamma}'w_q + \hat{\sigma}_{\xi}^2/2) - \exp(\hat{\alpha}'z_q + \hat{\sigma}_{\eta}^2/2)\right) \qquad \text{(B.1)}$$

$$\overset{\wedge}{\text{TT}} = \left[\frac{1}{Q_{r1}}\sum_{q=1}^{Q} r_q \times \left(\exp(\hat{b}_{q1} + \hat{\sigma}_{\xi}^2/2) - \exp(\hat{b}_{q0} + \hat{\sigma}_{\eta}^2/2)\right)\right] \qquad \text{(B.2)}$$

where Q_{r1} is the number of households in the sample residing in conventional neighborhoods, and $\hat{b}_{q0}$ and $\hat{b}_{q1}$ are defined as follows:

$$\hat{b}_{qo} = E(m_{qo} \mid r_q^* > 0) = \left\{1 - F_{\varepsilon}(-\hat{\beta}'x_q)\right\}^{-1} \times \frac{1}{\hat{\sigma}_{\eta}} \times \int_{m_{q0}} m_{qo} \times \left(1 - \frac{\partial C_{\theta_0}(u_{q1}^0, u_{q2}^0)}{\partial u_{q2}^0}\right) \times f_{\eta}\left(\frac{m_{q0} - \hat{\alpha}'z_q}{\hat{\sigma}_{\eta}}\right) dm_{qo},$$

$$\hat{b}_{q1} = E(m_{q1} \mid r_q^* > 0) = \left\{1 - F_{\varepsilon}(-\hat{\beta}'x_q)\right\}^{-1} \times \frac{1}{\hat{\sigma}_{\xi}} \times \int_{m_{q1}} m_{q1} \times \left(1 - \frac{\partial C_{\theta_1}(u_{q1}^1, u_{q2}^1)}{\partial u_{q2}^1}\right) \times f_{\eta}\left(\frac{m_{q1} - \hat{\gamma}'w_q}{\hat{\sigma}_{\xi}}\right) dm_{q1}.$$

The expressions above do not have a closed form in the general copula case. However, when a Gaussian copula is used for both the switching regimes, the expressions simplify nicely (see Lee, 1978). In the general copula case, the expressions (and the TT measure) can be computed using numerical integration techniques. It is also straightforward algebra to show that $\hat{b}_{q0} = \hat{\alpha}'z_q$ if there is no dependency in the (ε_q, η_q) terms, and $\hat{b}_{q1} = \hat{\gamma}'w_q$ if there is no dependency between the (ε_q, ξ_q) error terms. Thus, TT collapses to the ATE if the ATE were computed only across those households living in conventional neighborhoods (see the relationship between Equations (B.1) and (B.2) after letting $\hat{b}_{q0} = \hat{\alpha}'z_q$ and $\hat{b}_{q1} = \hat{\gamma}'w_q$ in the latter equation).

$$\overset{\wedge}{\text{TNT}} = \frac{1}{Q_{r0}}\left[\sum_{q=1}^{Q}(1 - r_q) \times \left(\exp(\hat{h}_{q1} + \hat{\sigma}_{\xi}^2/2) - \exp(\hat{h}_{q0} + \hat{\sigma}_{\eta}^2/2)\right)\right], \qquad \text{(B.3)}$$

where Q_{r0} is the number of households in the sample residing in neo-urbanist neighborhoods, and $\hat{h}_{q0}$ and $\hat{h}_{q1}$ are defined as follows:

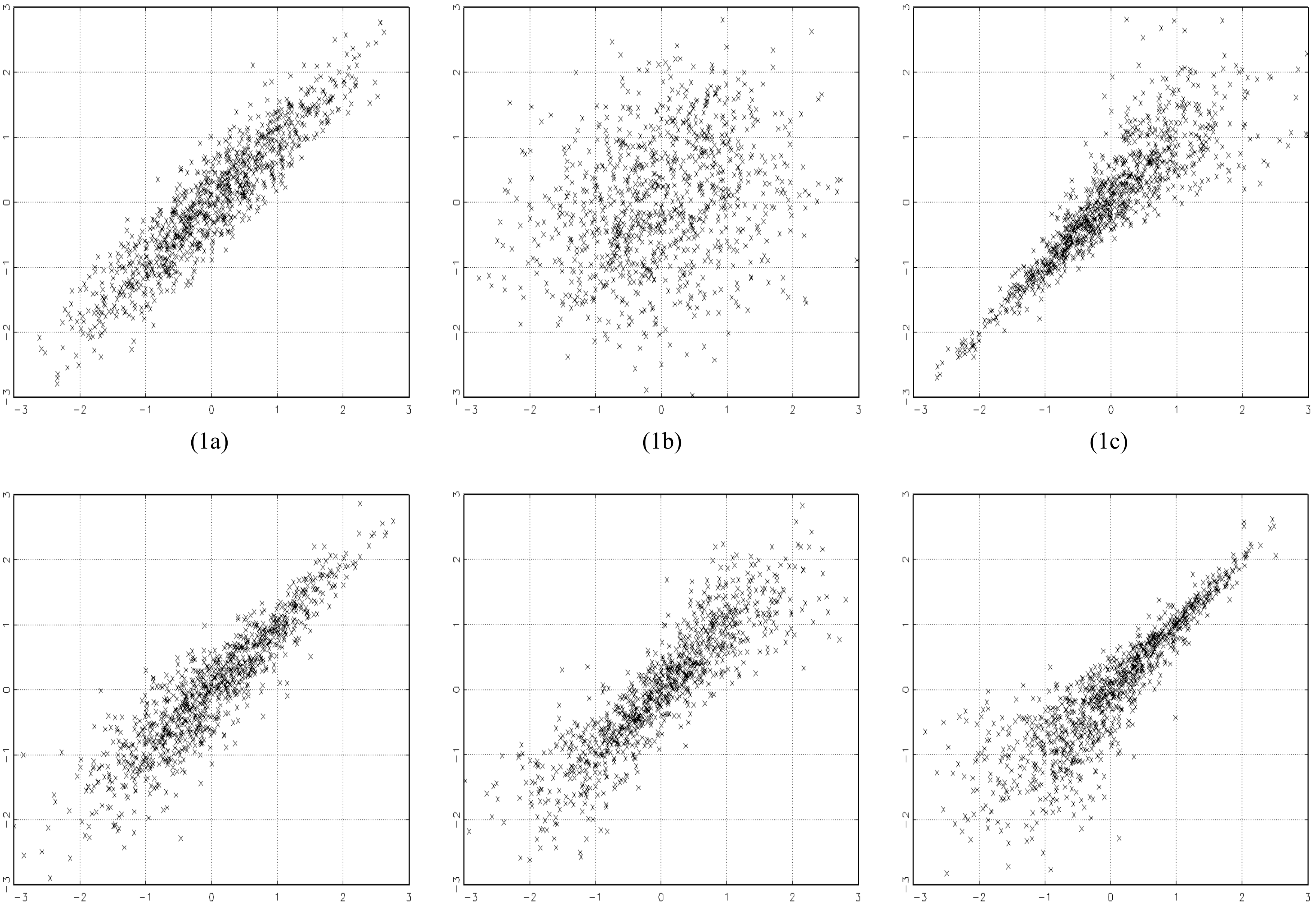
$$\hat{h}_{q0} = E(m_{q0} \mid r_q^* < 0) = \left\{F_{\varepsilon}(-\hat{\beta}'x_q)\right\}^{-1} \times \frac{1}{\hat{\sigma}_{\eta}} \times \int_{m_{q0}} m_{qo} \times \left(\frac{\partial C_{\theta_0}(u_{q1}^0, u_{q2}^0)}{\partial u_{q2}^0}\right) \times f_{\eta}\left(\frac{m_{q0} - \hat{\alpha}'z_q}{\hat{\sigma}_{\eta}}\right) dm_{qo},$$

$$\hat{h}_{q1} = E(m_{q1} \mid r_q^* < 0) = \left\{F_{\varepsilon}(-\hat{\beta}'x_q)\right\}^{-1} \times \frac{1}{\hat{\sigma}_{\xi}} \times \int_{m_{q1}} m_{q1} \times \left(\frac{\partial C_{\theta_1}(u_{q1}^1, u_{q2}^1)}{\partial u_{q2}^1}\right) \times f_{\eta}\left(\frac{m_{q1} - \hat{\gamma}'w_q}{\hat{\sigma}_{\xi}}\right) dm_{q1}.$$

$$\overset{\wedge}{\text{TTNT}} = \frac{1}{Q}\left(Q_{r0}\,\overset{\wedge}{\text{TNT}} + Q_{r1}\,\overset{\wedge}{\text{TT}}\right) \qquad \text{(B.4)}$$

LIST OF FIGURES

LIST OF TABLES



(1a) (1b) (1c)

(1d) (1e) (1f)

Figure 1 Normal variate copula plots (1a) Gaussian Copula τ = 0.75, θ = 0.92; (1b) FGM Copula τ = 0.22, θ = 1.00; (1c) Clayton Copula τ = 0.75, θ =6.00; (1d) Gumbel Copula τ = 0.75, θ = 4.00; (1e) Frank Copula τ = 0.75, θ = 14.14; (1f) Joe Copula τ = 0.75, θ = 6.79.

Table 1 Characteristics of Alternative Copula Structures

Copula	Dependence Structure Characteristics	Archimedean Generation Function $\psi(t)$	$\psi'(t)$	θ range and value for index	Kendall's τ and range	Spearman's ρ_S and range
Gaussian	Radially symmetric, weak tail dependencies, left and right tail dependencies go to zero at extremes	Not applicable	Not applicable	$-1 \le \theta \le 1$ $\theta = 0$ is independence	$\frac{2}{\pi}\arcsin(\theta)$ $-1 \le \tau \le 1$	$\frac{6}{\pi}\arcsin\left(\frac{\theta}{2}\right)$ $-1 \le \rho_S \le 1$
FGM	Radially symmetric, only moderate dependencies can be accommodated	Not applicable	Not applicable	$-1 \le \theta \le 1$ $\theta = 0$ is independence	$\frac{2}{9}\theta$ $-2/9 \le \tau \le 2/9$	$\frac{1}{3}\theta$ $-1/3 \le \rho_S \le 1/3$
Clayton	Radially asymmetric, strong left tail dependence and weak right tail dependence, right tail dependence goes to zero at right extreme	$\varphi(t) = \frac{1}{\theta}(t^{-\theta} - 1)$	$t^{-\theta-1}$	$0 < \theta < \infty$ $\theta \to 0$ is independence	$\frac{\theta}{\theta+2}$ $0 < \tau < 1$	No simple form $0 < \rho_S < 1$
Gumbel	Radially asymmetric, weak left tail dependence, strong right tail dependence, left tail dependence goes to zero at left extreme	$\varphi(t) = (-\ln t)^{\theta}$	$-\frac{\theta}{t}(-\log t)^{\theta-1}$	$1 \le \theta < \infty$ $\theta = 1$ is independence	$1-\frac{1}{\theta}$ $0 \le \tau < 1$	No simple form $0 \le \rho_S < 1$
Frank	Radially symmetric, very weak tail dependencies (even weaker than Gaussian), left and right tail dependencies go to zero at extremes	$\varphi(t) = -\ln\left[\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\right]$	$\frac{\theta}{1-e^{\theta t}}$	$-\infty < \theta < \infty$ $\theta \to 0$ is independence	See Equation (25) $-1 \le \tau \le 1$	$1-\frac{12}{\theta}(D_1 \mid \theta) - D_2(\theta)^{*}$ $-1 \le \rho_S \le 1$
Joe	Radially asymmetric, weak left tail dependence and very strong right tail dependence (stronger than Gumbel), left tail dependence goes to zero at left extreme	$\varphi(t) = -\ln[1-(1-t)^{\theta}]$	$\frac{-\theta(1-t)^{\theta-1}}{1-(1-t)^{\theta}}$	$1 \le \theta < \infty$ $\theta = 1$ is independence	See Equation (27) $0 \le \tau < 1$	No simple form $0 \le \rho_S < 1$

* $D_k(\theta) = \frac{k}{e^k}\int_{t=0}^{\theta} \frac{t^k}{(e^t - 1)} dt$

Table 2 Expressions for $\frac{\partial}{\partial u_2} C_\theta(u_1, u_2)$

Copula	**Expression**
Gaussian Copula	$\Phi\left[\frac{\Phi^{-1}(u_1) - \theta\,\Phi^{-1}(u_2)}{\sqrt{1-\theta^2}}\right]$
FGM Copula	$u_1[1+\theta(1-u_1)(1-2u_2)$
Clayton Copula	$u_2^{-(\theta+1)}(u_1^{-\theta} + u_2^{-\theta} - 1)^{-\left(\frac{1+\theta}{\theta}\right)}$
Gumbel Copula	$u_2^{-1}(-\ln u_2)^{\theta-1} \cdot C_\theta(u_1, u_2)\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{\left(\frac{1}{\theta}-1\right)}$
Frank Copula	$1 - e^{\theta u_2}(e^{\theta u_1} - e^\theta)\left[e^{\theta u_1} e^{\theta u_2} + e^\theta(1 - e^{\theta u_1} - e^{\theta u_2})\right]^{-1} = \left[1 - e^{\theta C_\theta(u_1, u_2)}\right](1 - e^{\theta u_2})^{-1}$
Joe Copula*	$\overline{u}_2^{\theta-1}(1-\overline{u}_1^\theta)\left[\overline{u}_1^\theta + \overline{u}_2^\theta - \overline{u}_1^\theta, \overline{u}_2^\theta\right]^{\left(\frac{1}{\theta}-1\right)}$

* For Joe's Copula, $\overline{u}_2 = 1 - u_2$, $\overline{u}_1 = 1 - u_1$

Table 3 Estimation Results of the Switching Regime Model

Variables	**Independence-Independence Copula**		**Frank-Frank Copula**	
	Parameter	**t-stat**	**Parameter**	**t-stat**
Propensity to choose conventional neighborhood relative to neo-urbanist neighborhood				
Constant	0.201	4.15	0.275	5.72
Age of householder < 35 years	-0.131	-2.35	-0.143	-2.75
Number of children (of age < 16 years) in the household	0.164	4.62	0.161	4.59
Household lives in a single family dwelling unit	0.382	6.79	0.337	6.28
Own household	0.597	10.37	0.497	8.81
Log of vehicle miles of travel in a neo-urbanist neighborhood				
Constant	-0.017	-0.16	-0.638	-5.48
Household vehicle ownership				
Household Vehicles = 1	2.617	21.50	2.744	24.26
Household Vehicles ≥ 2	3.525	25.44	3.518	27.40
Number of full-time students in the household	0.183	2.13	0.112	1.41
Copula dependency parameter (θ)	--	--	-2.472	-6.98
Scale parameter of the continuous component	1.301	40.62	1.348	34.31
Log of vehicle miles of travel in a conventional neighborhood				
Constant	0.379	2.28	0.163	1.08
Household vehicle ownership				
Household Vehicles = 1	3.172	21.77	3.257	25.43
Household Vehicles = 2	3.705	25.32	3.854	29.92
Household Vehicles ≥ 3	3.931	25.92	4.102	30.41
Number of employed individuals in the household	0.229	7.24	0.208	6.66
Number of full-time students in the household	0.104	5.06	0.131	6.27
Density of bicycle lanes	-0.023	-3.08	-0.024	-3.24
Accessibility to shopping (Hansen measure)	-0.024	-7.34	-0.027	-8.19
Copula dependency parameter (θ)	--	--	3.604	7.22
Scale parameter of the continuous component	0.891	75.78	0.920	63.59
Log-likelihood at convergence	**-6878.1**		**-6842.2**	

Table 4 Estimates of Treatment Effects in Miles

Copulas	Independence-Independence Copula (I-I)	FGM-Joe Copula (FG-J)	Frank-Joe Copula (F-J)	Frank-Frank Copula (F-F)
$\hat{ATE}$	0.49 (1.75)	10.75 (1.03)	19.99 (4.42)	21.37 (5.21)
$\hat{TT}$	3.04 (1.49)	31.04 (3.30)	42.45 (7.46)	41.76 (8.16)
$\hat{TNT}$	-8.38 (1.38)	-31.55 (10.06)	-33.66 (10.82)	-30.74 (9.55)
$\hat{TTNT}$	0.49 (1.75)	17.07 (0.88)	25.46 (3.03)	25.59 (4.75)