A Copula-based Method to Address Endogeneity in Traffic Crash Injury Severity Models: Application to Two-Vehicle Crashes

Tejsingh A. Rana
Department of Civil & Environmental Engineering
University of South Florida, ENC 1208
4202 E. Fowler Ave., Tampa, Fl 33620
Tel: (813) 507-1598; Fax: (813) 974-2957
Email: trana@mail.usf.edu

Sujan Sikder
Department of Civil & Environmental Engineering
University of South Florida, ENC 1208
4202 E. Fowler Ave., Tampa, Fl 33620
Tel: (813) 507-3953; Fax: (813) 974-2957
Email: ssikder@mail.usf.edu

Abdul Rawoof Pinjari (*corresponding author*)
Department of Civil & Environmental Engineering
University of South Florida, ENC 2503
4202 E. Fowler Ave., Tampa, Fl 33620
Tel: (813) 974- 9671; Fax: (813) 974-2957
Email: apinjari@eng.usf.edu

Submitted for Presentation and Publication to
Committee ANB20: Safety Data, Analysis and Evaluation

89th Annual Meeting of the Transportation Research Board
January 10-14, 2010

Word count: 7385 (text) + 4 (tables/figures) x 250 = 8385 equivalent words

Revised: November 2009

ABSTRACT

This paper employs a recently emerging copula-based methodology to address endogeneity in crash injury severity models. Specifically, two important sources of endogeneity are addressed: (1) Endogeneity due to the correlations between the injury severities of the two drivers involved in two-vehicle crashes, and (2) Endogeneity of collision type and injury severity outcomes. To this end, two sets of copula-based joint model systems are formulated and estimated using data on two-vehicle crashes from the 2007 Generalized Estimates System: (1) A copula-based joint ordered logit-ordered logit model of injury severities of the two drivers involved in two-vehicle crashes, and (2) A copula-based joint multinomial logit-ordered logit model of collision type and injury severity outcomes of two-vehicle crashes. To our knowledge, this study constitutes one of the first applications of the copula-based methods to address endogeneity in traffic safety literature.

Model estimation results from the two joint model systems show a statistically significant presence of the two types of endogeneity, ignoring which resulted in deteriorated model fit, biased parameter estimates and distorted elasticity/marginal effects. These results underscore the importance of accommodating endogeneity in crash injury severity modeling, as well as the potential of the copula-based methods in traffic crash modeling and analysis. Further, the model estimates offer useful insights on the impact of various roadway, environmental, vehicle, and driver characteristics on the injury severity of the drivers involved in two-vehicle accidents.

1 INTRODUCTION

Automobile crashes cause significant losses to the society due to fatalities, traffic congestion, medical costs, and property damage. The severity of injuries sustained by the individuals involved in these incidents depends on a multitude of factors, including roadway design features (alignment, pavement condition), environmental factors (weather, traffic conditions), vehicle characteristics (size, weight, safety features), and drivers' characteristics (age, gender) and their driving behavior (speed, seat belt use, alcohol influence, aggressiveness). It is important to understand and quantify the influence of each of these factors on crash injury severity so that engineering/policy measures can be formulated to reduce severe crashes. To this end, transportation safety literature abounds with studies [e.g., (*1*) and (*2*)] that model the relationship between crash injury severity and the factors identified above.

A major limitation of several crash injury severity studies in the literature, however, is the neglect of several sources of endogeneity bias. Even if endogeneity is recognized, the currently used methods to do so are associated with several drawbacks. Thus, this paper employs a new, copula-based methodology to address the issue of endogeneity in crash injury severity models. Specifically, two important sources of endogeneity are addressed in the context of two-vehicle crashes: (a) Endogeneity due to the correlations between the injury severities of the drivers involved in two-vehicle crashes, and (b) Endogeneity due to the common unobserved factors affecting the collision type and injury severity outcomes in two-vehicle crashes. To this end, we formulate and estimate: (a) a copula-based joint ordered logit-ordered logit (ORL-ORL) model system to jointly model the injury severity levels of the drivers involved in two-vehicle crashes, and (b) a copula-based joint multinomial logit-ordered logit (MNL-ORL) model system to jointly analyze the collision type and injury severity outcomes in two-vehicle crashes.

The next section explains the different sources of endogeneity in injury severity models and reviews the relevant literature. Section 3 provides an overview of copulas and formulates the copula-based joint model systems used in the paper. Section 4 presents the empirical analysis results using data from the 2007 Generalized Estimates System (GES) database. Section 5 concludes the paper.

2 ENDOGENEITY IN CRASH INJURY SEVERITY MODELING

Econometrically speaking, endogeneity bias occurs due to the presence of non-independent errors in the model specification (*3*). Such non-independent errors can occur due to several reasons, including: (a) The presence of endogenous explanatory variables where the observed explanatory variables in a model are correlated with the unobserved factors in the error terms, and (b) The presence of correlations across error terms of different model equations due to common unobserved factors influencing the outcome variables of interest.

A classic case of the former type of endogeneity in crash injury severity models corresponds to the use of seat belts. The decision to wear a seat belt is typically used as an exogenous explanatory variable to explain the influence of seat belt use on crash injury severity without considering the behavioral differences between those who wear seat belts and those who do not. However, it is possible that seat belt non-users may be intrinsically unsafe drivers [(*4*); (*5*)] and their unsafe driving habits may lead to severe crashes. The impact of such behavioral differences on injury severity outcomes, when not considered, may get confounded with the impact of seat belt use on injury severity outcomes. This econometric issue arises because the seat belt use variable in the model is likely to be correlated to the error term that contains unobserved factors (such as unsafe driving habits) influencing injury severity. In other words, common unobserved factors influence both seatbelt use and crash injury severity. Ignoring such endogeneity of an explanatory variable can lead to biased and inconsistent estimates and distorted policy implications (*3*).

Similar to the endogeneity of seat belt-use, several other factors such as crash type (head-on, angle, rear-end, etc.), driving under alcohol/drug influence, driving during night time, and vehicle occupancy may be endogenous to the severity of crashes. For example, in the context of crash type, it is very well recognized that head-on collisions are likely to result in more severe injuries than other types (angle, rear-end, and side-swipe) of collisions involving multiple vehicles. To accommodate this, collision type is typically included as an exogenous explanatory variable. However, there may be several unobserved driver characteristics or roadway features that may result in a particular type(s) of collision(s) as well as influence the severity of injuries. Due to such influence of common unobserved factors, the collision type variable(s) can potentially be endogenous to injury severity outcomes.

Another important type of endogeneity occurs due to common unobserved factors affecting the injury severity of different individuals involved in a same crash. For example, if a driver involved in a two-vehicle crash sustains severe injuries, it is likely that the other driver (and other individuals) involved in that crash also sustains severe injuries. This is because several factors (such as speeding/aggressive driving and roadway design/environment-related factors) that influence the injury severity of one driver also influence the injury severity of another driver. Thus, the injury severity propensities of all the individuals involved in a crash can potentially be correlated to (or endogenous to) each other. Despite the likelihood of such endogeneity, most crash injury severity studies in the literature model the injury severity of each crash victim as independent of the injury severity of other victims from the same crash as if each individual was involved in a different crash.

To be sure, the traffic safety literature is not devoid of studies that recognize and capture the forms of endogeneity discussed above. Table 1 provides an extensive coverage and summary of these studies. The first six studies in the table focus on the endogeneity of seat belt-use. Among these, while Evans (*4*) addressed this issue using simple descriptive analysis, Dee (*6*), Derrig *et al.* (*7*) and Cohen and Einav (*8*) used aggregate-level analyses using data on seat belt usage rates and fatality rates. Eluru and Bhat (*5*) recognized that such aggregate-level analyses mask the heterogeneity across individual crashes and used a disaggregate crash victim-level injury severity analysis to address the issue of seat belt endogeneity. They propose a joint random coefficients modeling approach where common random terms are employed across the seat belt use and injury severity models to capture the endogeneity due to common unobserved factors affecting seat belt use and injury severity. To do the same, de Lapparent (*9*) used the familiar, bivariate probit modeling approach.

Only two studies have endogenously analyzed crash type with injury severity. Of these two studies, while Kim *et al.* (*10*) include crash type as an endogenous variable in a structural equations framework, Ye *et al.* (*11*) explicitly recognize the endogeneity of two-vehicle collision type variables using a joint, random coefficients multinomial logit – ordered logit model of collision type and injury severity. Their results indicate a significant presence of unobserved factors affecting crash type and injury severity, ignoring which may result in inconsistent and biased coefficient estimates in the injury severity model.

Lee and Abdel-Aty (*12*) used bivariate probit models to address the endogeneity of vehicle occupancy and passenger characteristics in crash outcome (injury severity, type, etc.) models. Such endogeneity of explanatory variables has been found to be an important issue not only in crash injury severity models, but also in crash frequency models. For example, Kim and Washington (*13*) identified that ignoring the endogeneity of left-turn lane variables may lead to a counterintuitive result that left-turn lanes cause an increased occurrence of angle accidents.

The next set of studies [(*14*) - (*17*)] addresses the endogeneity due to the simultaneity of the injury severity outcomes of the different individuals involved in a same crash. These studies use bivariate ordered (or binary) response modeling methods to jointly model the injury severities of the two individuals under consideration. The next set of studies [(*18*) - (*21*)], are further advanced in that they recognize the presence of a multilevel hierarchy (such as vehicle-level, crash-level, and location-level) in traffic crash data. These studies provide evidence of significant magnitude of common unobserved factors at each level of the hierarchy. Thus, for example, the unobserved factors at the crash-level cause correlations among the injury severity model equations of all individuals involved in the crash.

Each of the methods used in the above-discussed studies is associated with specific drawbacks. For example, the joint random coefficients modeling approach necessitates the use of simulation-based estimation, which is computationally intensive and saddled with such technical issues as bias and unidentifiability of model parameters. The bivariate probit method is not easily adaptable to accommodate the endogeneity of polychotomous categorical variables (such as collision type). Another disadvantage is that the bivariate normal distribution used for most bivariate modeling methods is restrictive in the types of correlations (or dependencies) it can accommodate between the variables of interest (more in Section 3). Further, to extend this approach to model multivariate outcomes, one has to resort to computationally intensive simulation-based estimation. The multi-level modeling method, as identified by Lengeurrand *et al.* (*19*) is better applicable "*when the number of vehicles per crash and the number of occupants per vehicle is high*". Their simulation experiments indicated that when the number of data points in a cluster (or level) is small; such models may be difficult to estimate. In the context of crash injury severity modeling, since a large proportion of crashes do not involve more than two individuals per vehicle and more than two vehicles per crash, an alternative approach may be preferred. Further, even if estimable, multi-level models require the use of computationally intensive simulation-based estimation methods.

In view of the disadvantages of the currently used methods, this paper uses a recently emerging copula-based approach to address endogeneity in crash injury severity models for drivers involved in two-vehicle crashes. Specifically, two important sources of endogeneity are addressed in this paper: (a) Endogeneity due to common unobserved factors influencing the injury severity of the two drivers involved in a crash, and (b) Endogeneity of collision type outcomes to injury severity outcomes. The copula-based approach, as discussed next, offers several advantages over other methods used in the literature.

3 COPULA-BASED METHODOLOGY

Copulas are mathematical constructs used to generate dependency (or correlations) among stochastic variables with known marginal distributions [(*22*); (*23*)]. Specifically, a copula is a multivariate distribution function defined to link (or tie) several uniformly distributed marginal variables. Following Bhat and Eluru (*23*), if U_1, U_2,…,U_N are N uniformly distributed random variables, then the N-dimensional copula or the N-dimensional joint distribution of these random variables can be defined as:

$$C_\theta (u_1, u_2, \ldots, u_N) = \Pr(U_1 < u_1, U_2 < u_2, \ldots, U_N < u_N), \quad (1)$$

where θ is a parameter vector referred to as the dependence parameter vector analogous to (but not the same as) the correlation vector in a multivariate normal distribution. Such a copula function can be applied to pre-specified marginal distributions to generate dependency (or correlations) among those marginal distributions. To see this, again following Bhat and Eluru (*23*), consider N univariate random variables X_1, X_2,…,X_N, each with continuous marginal distribution functions $F_n(y_n) = \Pr(Y_n < y_n)$, n =1,2,…,N. Using the integral transform result, the marginal distribution of each random variable X_n can be expressed as:

$$F_n(x_n) = \Pr(X_n < x_n) = \Pr(F_n^{-1}(U_n) < x_n) = \Pr(U_n < F_n(x_n)). \tag{2}$$

Then, by Sklar's (*24*) theorem, a joint N-dimensional distribution function of the random variables with the marginal distribution functions $F_n(x_n)$ can be generated as:

$$\begin{aligned} F(x_1, x_2, \ldots, x_n) &= \Pr(X_1 < x_1, X_2 < x_2, \ldots, X_n < x_n) \\ &= \Pr(U_1 < F_1(x_1),, U_2 < F_2(x_2), \ldots, U_n < F_n(x_n)) \end{aligned} \tag{3}$$

which is nothing but a copula function (as in Equation 1) as below:

$$F(x_1, x_2, \ldots, x_n) = C_\theta\,(u_1 = F_1(x_1), u_2 = F_2(x_2), \ldots, u_n = F_n(x_N)). \tag{4}$$

There are several advantages to using copulas. First, copulas can capture more general forms of dependency (such as asymmetric dependence, and asymptotic dependence) than the simple, symmetric and asymptotically independent forms of dependency exhibited in bivariate normal distributions of the bivariate probit models. Second, the dependency form is independent of the marginal distributions of the stochastic variables. Thus the stochastic variables of interest (for example, injury severity propensity and collision type propensity) need not necessarily follow the same marginal distribution. Third, a variety of copula functions can be used to explore different forms of dependency as opposed to the usual bivariate normal distributions used in earlier studies. Fourth, several copulas offer closed-form probability expressions and obviate the need for computationally intensive simulation-based model estimation.

In the context of crash injury severity modeling, copulas can be used to join the injury severity models of different individuals involved in an accident into a simultaneous equations framework. Further, copulas can also be applied to incorporate endogeneity of collision type (or other) variables. In this section, we first formulate the familiar ordered logit model of injury severity, then develop two copula-based joint model systems: (a) A copula-based joint ordered logit-ordered logit (ORL-ORL) structure to jointly model the injury severity levels of two drivers involved in a two-vehicle crash, and (b) A copula-based multinomial logit-ordered logit (MNL-ORL) structure to jointly model collision type and driver injury severity.

3.1 Independent Injury Severity Model

Let q_{dj} be an index to represent the two drivers d (d = 1,2) involved in a two-vehicle collision/accident q (q = 1,2,…,Q) of type $j\ (j = 1, 2, \ldots J)$ and let k_{dj} (k_{dj} = 1,2,3,…,K) be an index to represent injury severity of the drivers. In the current paper, j takes the values of "head-on" (j =1), "angle" (j =2), "rear-end" (j =3), "sideswipe" (j =4) and "other" (j =5), where the index k_{dj}, takes the values of "no injury" (k_{dj} = 1), "possible injury" (k_{dj} = 2), "non-incapacitating injury" (k_{dj} = 3), "incapacitating injury" (k_{dj} = 4), and "fatal injury" (k_{dj} = 5). Let $y_{q_{dj}}$ denote the observed injury severity sustained by the drivers involved in a two-vehicle collision, let $y^*_{q_{dj}}$ denote the latent (unobserved) injury severity propensity of those drivers, and let $\psi_{k_{dj}}$ be the thresholds used to map the observed injury severity levels to the latent injury severity propensities. Using this notation, define $y^*_{q_{dj}}$ as below:

$$y^*_{q_{dj}} = \beta'_j x_{q_{dj}} + \xi_{q_{dj}} \tag{5}$$

where β'_j is a vector of coefficients of the observed roadway, environmental, vehicle, and driver characteristics $x_{q_{dj}}$ affecting the driver's injury severity if (s)he was involved in an accident of type j, and $\xi_{q_{dj}}$ is a random component that captures the unobserved factors affecting the injury severity. The latent injury severity propensity $y^*_{q_{dj}}$ of each driver d is mapped to his/her injury severity level $y_{q_{dj}}$ by the $\psi_{k_{dj}}$ thresholds as below:

$$\begin{aligned}
y_{q_{dj}} &= k_{dj} \text{ if } \psi_{k_{dj}-1} < y^*_{q_{dj}} < \psi_{k_{dj}}, or \\
y_{q_{dj}} &= k_{dj} \text{ if } \psi_{k_{dj}-1} < \beta'_j x_{q_{dj}} + \xi_{q_{dj}} < \psi_{k_{dj}}, or \\
y_{q_{dj}} &= k_{dj} \text{ if } (\psi_{k_{dj}-1} - \beta'_j x_{q_{dj}}) < \xi_{q_{dj}} < (\psi_{k_{dj}} - \beta'_j x_{q_{dj}})
\end{aligned} \tag{6}$$

Given the above mapping, assuming that the $\xi_{q_{dj}}$ terms are Gumbel distributed, the probability of a driver d involved in a two-vehicle accident q of type j sustaining an injury severity of level k_{dj} ($y_{q_{dj}} = k_{dj}$) is given by the familiar ordered logit formula:

$$P(y_{q_{dj}} = k_{dj}) = F_{\xi_d j}(\psi_{k_{dj}} - \beta'_j x_{q_{dj}}) - F_{\xi_d j}(\psi_{k_{dj}-1} - \beta'_j x_{q_{dj}}) \tag{7}$$

where, $F_{\xi_d j}(.)$ is the cumulative distribution function of the error term $\xi_{q_{dj}}$. This probability expression represents an independent injury severity model (for each driver and accident type j) that does not capture any form of endogeneity.

3.2 Joint Injury Severity Model for the Two Drivers Involved in Two-Vehicle Crashes

The equation system for simultaneously modeling the injury severity of two drivers q_{1j} and q_{2j} involved in a two-vehicle accident q of type j can be written as:

$$\begin{aligned}
y_{q_{1j}} &= k_{1j} \text{ if } (\psi_{k_{1j}-1} - \beta'_j x_{q_{1j}}) < \xi_{q_{1j}} < (\psi_{k_{1j}} - \beta'_j x_{q_{1j}}) \\
y_{q_{2j}} &= k_{2j} \text{ if } (\psi_{k_{2j}-1} - \beta'_j x_{q_{2j}}) < \xi_{q_{2j}} < (\psi_{k_{2j}} - \beta'_j x_{q_{2j}})
\end{aligned} \tag{8}$$

From the above equation system, the joint probability that one driver sustains injuries of severity level k_{1j} and another driver sustains injuries of severity of level k_{2j} is:

$$
\begin{aligned}
& P(y_{q_{1j}} = k_{1j}, y_{q_{2j}} = k_{2j}) = \\
& P\left([(\psi_{k_{1j}-1} - \beta_j' x_{q_{1j}}) < \xi_{q_{1j}} < (\psi_{k_{1j}} - \beta_j' x_{q_{1j}})],\ [(\psi_{k_{2j}-1} - \beta_j' x_{q_{2j}}) < \xi_{q_{2j}} < (\psi_{k_{2j}} - \beta_j' x_{q_{2j}})] \right) \\
& = \ P\,[\xi_{q_{1j}} < (\psi_{k_{1j}} - \beta_j' x_{q_{1j}}),\ \ \xi_{q_{2j}} < (\psi_{k_{2j}} - \beta_j' x_{q_{2j}})] \\
& \quad - P\,[\xi_{q_{1j}} < (\psi_{k_{1j}} - \beta_j' x_{q_{1j}}),\ \ \xi_{q_{2j}} < (\psi_{k_{2j}-1} - \beta_j' x_{q_{2j}})] \\
& \quad - P\,[\xi_{q_{1j}} < (\psi_{k_{1j}-1} - \beta_j' x_{q_{1j}}),\ \ \xi_{q_{2j}} < (\psi_{k_{2j}} - \beta_j' x_{q_{2j}})] \\
& \quad + P\,[\xi_{q_{1j}} < (\psi_{k_{1j}-1} - \beta_j' x_{q_{1j}}),\ \ \xi_{q_{2j}} < (\psi_{k_{2j}-1} - \beta_j' x_{q_{2j}})]
\end{aligned}
\tag{9}
$$

The form of the above probability expression depends on the specification of the dependency form between the random terms $\xi_{q_{1j}}$ and $\xi_{q_{2j}}$. Specifically, one may use copula functions to write the joint probability expression as:

$$
P(y_{q_{1j}} = k_{1j}, y_{q_{2j}} = k_{2j}) = C_\theta(u^j_{qk_1}, u^j_{qk_2}) - C_\theta(u^j_{qk_1}, u^j_{qk_2-1}) - C_\theta(u^j_{qk_1-1}, u^j_{qk_2}) + C_\theta(u^j_{qk_1-1}, u^j_{qk_2-1}) \tag{10}
$$

where, $C_\theta(.\,,\,.)$ is a copula function defining the dependency form between $\xi_{q_{1j}}$ and $\xi_{q_{2j}}$, and

$$
\begin{aligned}
& u^j_{qk_1} = F_{\xi_1 j}\ (\psi_{k_{1j}} - \beta_j' x_{q_{1j}}),\ \ u^j_{qk_1-1} = F_{\xi_1 j}\ (\psi_{k_{1j}-1} - \beta_j' x_{q_{1j}}), \\
& u^j_{qk_2} = F_{\xi_2 j}\ (\psi_{k_{2j}} - \beta_j' x_{q_{2j}}),\ \ u^j_{qk_2-1} = F_{\xi_2 j}\ (\psi_{k_{2j}-1} - \beta_j' x_{q_{2j}}).
\end{aligned}
$$

Using the joint probability expression of Equation (10) for the injury severity of the two drivers involved in a two-vehicle accident *q*, the likelihood function of injury severity outcomes for all the Q two-vehicle accidents is:

$$
L = \prod_{q=1}^{Q} \left\{ \prod_{k_{1j},\, k_{2j}=1}^{K} \left[P(y_{q_{1j}} = k_{1j}, y_{q_{2j}} = k_{2j}) \right]^{\delta_{qk_{1j}} \delta_{qk_{2j}}} \right\}^{w_q} \tag{11}
$$

where $\delta_{qk_{1j}}$ and $\delta_{qk_{2j}}$ are dummy variables taking the value 1 if driver 1 and driver 2 involved in accident *q* of type *j* sustain injuries of levels k_{1j} and k_{2j}, respectively, and 0 otherwise. w_q is the weight for accident *q* used to represent an unbiased sample of two-vehicle crashes. The logarithm of the above likelihood function is maximized (i.e., the maximum likelihood approach is adopted) to estimate all parameters of the models, including the vector of coefficients β, the thresholds, and the dependency parameter θ. Maxlik module of the GAUSS matrix programming language was used to code the likelihood function and estimate the parameters.

3.3 Joint Model of Collision Type and Injury Severity

In the joint model of collision type and injury severity, the injury severity model component takes the ordered logit specification as in Equations (6) and (7) and the collision type model component takes the familiar discrete choice formulation. Consider the following equation that represents the propensity of an accident type j:

$$
u^*_{qj} = \alpha'_j z_{qj} + \varepsilon_{qj} \tag{12}
$$

In the equation above, u^*_{qj} is the propensity that q^{th} accident is of type $j\,(j = 1, 2, \ldots J)$, z_{qk} is a column vector of roadway design and environment, vehicle, and other attributes (including a constant) affecting the propensity, α'_j is the corresponding coefficient vector, and ε_{qj} is the error term capturing the effects of unobserved factors on the propensity associated with accident type j. With this propensity specification, the accident type outcome of an accident q is assumed to be of type j if it is associated with the maximum propensity among all J accident types; that is, if

$$u^*_{qj} > \max_{l=1,2,\ldots,J,\, l \neq J} u^*_{ql} \tag{13}$$

Next, following Lee (*25*) and Spissu *et al.* (*26*), the above polychotomous outcome model is recast into of a series of binary outcome model formulations, one for each collision type. To do so, let R_{qj} be a binary variable that takes a value of 1 if accident q is of type j and 0 otherwise. Subsequently, substituting $\alpha'_j z_{qj} + \varepsilon_{qj}$ for u^*_{qj} (from Equation 12), rewrite Equation (13) as:

$$R_{qj} = 1 \text{ if } \alpha'_j z_{qj} > v_{qj},\ (j = 1, 2, \ldots J) \tag{14}$$

$$\text{where, } v_{qj} = \{\max_{l=1,2,\ldots,J,\, l \neq j} u^*_{ql}\} - \varepsilon_{qj} \tag{15}$$

Equation (14) represents a series of binary outcome model formulations one for each collision type j, which is equivalent to the multinomial discrete choice model of collision type. An assumption that the ε_{qj} terms are assumed to be independent (across *j*) and identical Gumbel distributed results in logistic distribution for the v_{qj} terms, and consequently, the collision type probability expressions resemble the multinomial logit probabilities.

Now, the joint probability that an individual gets involved in a collision of type *j*, and sustains injuries of severity level k_{dj} is given by:

$$\begin{aligned} P(R_{qj} = 1, y_{q_{dj}} = k_{dj}) &= P\left\{\left(\alpha'_j z_{qj} > v_{qj}\right), \left((\psi_{k_{dj}-1} - \beta'_j x_{q_{dj}}) < \xi_{q_{dj}} < (\psi_{k_{dj}} - \beta'_j x_{q_{dj}})\right)\right\} \\ &= P\left((\alpha'_j z_{qj} > v_{qj}), (\xi_{q_{dj}} < \psi_{k_{dj}} - \beta'_j x_{q_{dj}})\right) - P\left((\alpha'_j z_{qj} > v_{qj}), (\xi_{q_{dj}} < \psi_{k_{dj}-1} - \beta'_j x_{q_{dj}})\right) \end{aligned} \tag{16}$$

The above joint probability expression depends upon the dependence structure between the random variables v_{qj} and $\xi_{q_{dj}}$. As indicated earlier, in this paper, copula-based methods are used to capture these dependences. To do so, first the marginal distributions of v_{qj} and $\xi_{q_{dj}}$, $F_{vj}(.)$ and $F_{\xi_d j}(.)$, are transformed into uniform distributions using their inverse cumulative distribution functions. Subsequently, copula functions are applied to "couple" the marginal inverse cumulative distribution functions into a joint distribution $F_{vj,\xi_d j}(.,.)$. Thus, for a driver involved in an accident *q*, the joint probability that the collision type outcome is *j* and the injury severity outcome is k_{dj} can be expressed using copula functions as below:

$$P(R_{qj} = 1, y_{q_{dj}} = k_{dj}) = \left[C_\theta(u^j_{qk_d}, u^j_q) - C_\theta(u^j_{qk_d-1}, u^j_q)\right] \tag{17}$$

where $C_{\theta j}(.,.)$ is the copula corresponding to $F_{vj,\xi_d j}(u^j_{qk_d}, u^j_q)$ and $F_{vj,\xi_d j}(u^j_{qk_d -1}, u^j_q)$, with $u^j_{qk_d} = F_{\xi_d j}(\psi_{k_{dj}} - \beta'_j x_{q_{dj}})$, $u^j_{qk_d -1} = F_{\xi_d j}(\psi_{k_{dj}-1} - \beta'_j x_{q_{dj}})$ and $u^j_q = F_{vj}(\alpha'_j z_{qj})$. This copula function captures the dependency between $v_{qj}\, (= \{\max_{l=1,2,\ldots,J, l\neq j} u^*_{ql}\} - \varepsilon_{qj})$ and $\xi_{q_{dj}}$ terms.

Using the joint probability expression of Equation (17) for the collision type and injury severity outcomes, the likelihood function for all the Q two-vehicle accidents is:

$$L = \prod_{q=1}^{Q} \left[\prod_{j=1}^{J} \prod_{k_{dj}=1}^{K} \left\{ P(R_{qj}=1, y_{q_{dj}} = k_{dj}) \right\}^{R_{qj}\delta_{qk_{dj}}} \right]^{w_q} \tag{18}$$

where $\delta_{qk_{dj}}$ takes the value 1 if driver *d* involved in accident *q* sustain injury of level k_{dj} and 0 otherwise (R_{q_j} and w_q re as defined earlier). The logarithm of the above likelihood function is used in a maximum likelihood estimation routine (coded in GAUSS) to estimate the model parameters

4 EMPIRICAL ANALYSIS

4.1 Data

The data used in this study was obtained from the 2007 Generalized Estimates System (GES) data made available by the National Highway Traffic Safety Administration. The 2007 GES includes information regarding 61,282 crashes (of which 35.7% were single vehicle crashes, and 55.9% were two-vehicle crashes) involving about 152,727 individuals and 107,202 vehicles.

From the GES data pool, information pertaining to two-vehicle crashes (consisting of 57,550 drivers involved in 28,775 two-vehicle crashes) was extracted. Further, crashes involving commercial vehicles or large trucks were discarded and the remaining data with missing information was removed. Subsequently, 5027 crash records (with 10054 driver records) were randomly sampled for model estimation and empirical analysis. Accident level weights were developed such that the weighted distribution of injury severity in this sample was the same as that in the full sample of two-vehicle drivers.

Among the 10054 driver records in the sample, close to 69.9% of the drivers experienced no injury, 14.3% driver records indicate possible injury, 9.1% indicate non-incapacitating injury, 6.1% indicate incapacitating injury, and 0.6% indicate fatal injury. Among the 5027 two-vehicle crashes, 6.2% are head-on collisions, 46.4% are angle collisions, 38.5% are rear-end collisions, 7.6% are sideswipe crashes, and 1.2% are other types of crashes. Further details and descriptive analyses (available from the authors) are being suppressed due to space considerations.

4.2 Joint Injury Severity Model for the Two Drivers Involved in Two-Vehicle Crashes

This section presents the results of the bivariate copula-based joint ORL-ORL model (of the injury severity of the two drives involved two-vehicle crashes) as well as a simple independent ORL model (that does not consider correlations between the injury severities of the two drivers). Table 2 presents the parameter estimates (and the corresponding t-statistics in parenthesis) from both the models (see second and third columns). Further, the marginal effects of each variable (computed using both the models) are also provided for the fatal injury category (see fourth and fifth columns).

Various driver, roadway, vehicle, and environmental attributes were explored in the model specification. The first set of variables in table corresponds to driver characteristics. Among these, the female dummy variable coefficients and marginal effects indicate that females are more susceptible to higher injury severities than males. Age-related variable effects indicate that drivers of age less than 65 years are less prone toward higher severity injuries than those of 65 years or older age. Alcohol/drug influence variable effects indicate that drivers under the influence of alcohol or drugs are likely to experience higher injury severities than those who do not drive under the influence of alcohol or drugs. Further, the coefficients and the marginal effects of the alcohol/drug use by the driver of the partner vehicle indicates that in two-vehicle accidents, even if a driver is not under the influence of alcohol or drugs, he/she is likely to experience higher injury severities if the other driver is involved in alcohol or drugs. This result indicates that those who drive under the influence of alcohol or drugs pose a risk of higher injury severity (and fatality) not only to themselves but also to other individuals involved in the accident. Although a very intuitive finding, not many previous studies have documented this result. Finally, among the driver characteristics, the seat belt effects indicate that use of safety belts helps in protecting the occupants from severe injuries.

Among the roadway characteristics, roadway surface condition related effects indicate that accidents occurring during snow or ice on the road tend to be less severe than those occurring during dry or wet conditions. This may be because drivers may exercise higher caution during such roadway conditions than normal (dry/wet) conditions. Similarly, drivers appear to be more cautious on steeper roads than on level roads. The next, speed limit variable effects indicate that driver injury tends to be most severe for crashes on medium speed limit (26-65mph) and high speed limit (>65mph) roads, when compared to low speed limit roads. Since vehicle speeds are higher on medium speed limit roads and high speed limit roads, the injury severity on such roads is higher than that on low speed limit roads. Since speed limit variables can be viewed as surrogates for (and are highly correlated with) road type classification (freeways, arterials, local streets, etc), we did not include road type classification variables in the model.

Crashes occurring in dark conditions tend to be more severe than those occurring during daylight. Land-use variable effects indicate that crashes occurring in areas of larger population (>100,000) tend to be more severe. The reason behind this result is not clear as one would expect areas with larger population to be associated with lower traffic speeds (due to high traffic volumes and congestion) and hence lower injury severity. The manner of collision variable effects indicate that head-on and angle collisions lead to more severe injuries, and side swipe collisions lead to less severe injuries than rear-end and other types of collisions (see Ye *et al.* (*11*) for a similar finding). The marginal effects indicate that head-on collisions show the largest propensity to cause fatalities. Vehicle role variable effects suggest a higher injury severity level if the vehicle is struck, or is struck and strikes another vehicle, relative to striking another vehicle.

Among the body type variables, as expected, drivers in pickup trucks and utility vehicles involved in two-vehicle accidents appear to be less prone to higher injuries than those in sedans. Further, if the partner (or other) vehicle involved in the accident is a non-sedan (*i.e.,* a pickup truck, or a utility vehicle), the driver injury severity tends to be higher (Kockelman and Kweon (*27*) report a similar result). Drivers of older vehicles (older than 10 years) indicate higher injury severity propensity than those of new vehicles, perhaps due to the improved safety features in new vehicles.

The next set of parameters corresponds to the thresholds, all of which, as expected, are in ascending order according to the levels of injury severity they correspond to.

The next, copula dependency parameter θ (for the copula-based joint model) represents the level of association (or correlation) between the injury severity propensities of the two drivers in two-vehicle accidents. In this model, we explored different types of copula functions (including Gaussian, Frank, Gumbel, Joe, and Clayton copula functions) to model the association between the two injury severity propensities. The Gumbel copula provided the best model fit. According to the properties of the Gumbel copula, θ value greater than 1 indicates a non-zero correlation. The reported t-statistic of the parameter (against a null value of 1) shows that the θ parameter value of 1.31 is statistically different from 1, indicating significant positive correlations between the injury severities of the two drivers involved in a same accident (see Hutchinson (*14*); and Ouyang et al. (*16*) for a similar result).

The positive correlation indicates that the unobserved factors that increase the injury severity of one driver involved in a two-vehicle accident also increase the injury severity of the other driver involved in that accident. Such correlations may arise because of the presence of several common (to the injury severity propensity equations of the two drivers) unobserved, but influential, factors that affect the injury severity of both the drivers involved in two-vehicle accidents. Such factors include, for example, vehicle speeds, aggressive/risky driving behavior habits (such as over speeding), and other roadway related features (such as presence or absence of guard rails, traffic conditions, etc.) that are not usually well-recorded in crash reports.

The log-likelihood ratio statistic between the independent and the copula-based joint models is -2*(-4747 – -4923) = 352, which is much higher than that the critical chi-square value for a degree of freedom of 1 (for one additional copula parameter in the joint model) at any level of significance. This indicates the superior statistical fit of the joint model than the independent model and that driver injury severity should be modeled in a joint fashion for two-vehicle crashes. The independent injury severity model treats the two drivers involved in a two-vehicle accident as if from two separate accidents. Such assumption may result in distorted estimates of the influence of various roadway, environmental, vehicle, and driver characteristics on injury severity. This can be observed by comparing the model estimates of the independent and copula-based joint models in Table 2. Further, the differences in the marginal effects obtained from the two models are non-negligible for certain variables. Specifically, the joint model shows smaller marginal effects than the independent model. For example, the magnitude of the marginal effect of seat belt use variable from the joint model is considerably lower than that from the independent model. This may be because safety belt non-users may be intrinsically unsafe drivers (*5*), because of which both the drivers in a two-vehicle accident are likely to experience injuries of high severity (especially fatal injuries). The joint model captures such common unobserved factors that affect the injury severity of both the drivers in the copula dependency parameter, and isolates the seat belt effect from any such confounding effects. Thus, the model indirectly controls for the endogeneity of the seat belt variable. On the other hand, the independent model simply ignores such unobserved factors that get confounded into the effect of seat belt use and lead to spuriously inflated estimates of the influence of seat belt use. Similar inferences can be made in the context of the differences in marginal effects of the head-on collision variable and the vehicle role (i.e., “struck by and strikes other vehicle”) variable. Thus, an advantage of jointly modeling the injury severities of the different individuals involved in a same crash is that different sources of endogeneity can be controlled for.

4.3 Results of the Joint Model of Collision Type and Injury Severity

The discussion in the preceding section corresponds to the endogeneity due to the correlations between the unobserved factors affecting the injury severities of different drivers. This section, presents the results of the copula-based joint MNL-ORL model of collision type and injury severity of drivers involved in two-vehicle collisions that accounts for the endogeneity of collision type. Table 3 presents the parameter estimates of the model. Due to space considerations, we briefly discuss the model estimates and focus on salient aspects.

The estimates of the collision type model component are in the left half of Table 3. These estimates indicate the influence of various roadway, crash and driver characteristics on the collision type outcome of a two-vehicle crash (given the occurrence of a two-vehicle crash). For example, a two-vehicle crash is likely to be a head-on collision when people drive under the influence of alcohol/drugs, in snow/ice road conditions, on road segments that are either curved or have no median, and during dark conditions. Similarly, a two-vehicle crash is likely to be an angle collision when it occurs on straight road segments (as opposed to curved segments), roads with no median, roads with multiple lanes, during dark conditions, in areas with large populations (>100,000), at intersections, and at roadway sections with stop signs and yield signs. Further, two-vehicle crashes that occur on snow/ice road conditions, steep road segments, at intersections, and at roadway sections with yield signs and traffic signals are likely to be rear-end collisions. Whereas the two-vehicle crashes that occur during snow/ice conditions, on roads with multiple lanes, during dark conditions, and in areas with large populations (>100,000) are likely to be sideswipes.

The estimates of the injury severity component are in the right half of the table. As can be observed, two copula dependency parameters (θ_j) were estimated, one each for head-on and angle crashes (The dependency parameters for rear-end, sideswipe and other types of crashes were not statistically significant). For both these types of crashes, the copula functions corresponding to Frank copulas provided the best model fit. For Frank copula functions, a dependency parameter significantly different from zero indicates a significant dependency (or correlation) between the marginal variables of interest [see Bhat and Eluru, (*23*)]. In this context, the implication is that there is significant positive correlation between the $v_{qj} (= \{\max_{l=1,2,\ldots,J,l\neq j} u^*_{ql}\} - \varepsilon_{qj})$ and $\xi_{q_{dj}}$ terms for head-on and angle collisions (see Section 3.3). This implies that the implied correlation between collision type propensity error term (ε_j) and injury severity propensity error term ($\xi_{q_{dj}}$) is negative and statistically significant. In the context of interpreting this result, it is important to note that the negative sign of the correlation term does not necessarily imply that injury severity in head-on/angle collisions is likely to be lower. Rather, the negative correlation indicates that the unobserved factors that contribute to the likelihood of head-on and angle collisions (within two-vehicle crashes) are negatively correlated with the unobserved factors that contribute to higher injury severity [see Ye *et al.* (*11*) for a similar result on the sign of correlations]. Further investigation is warranted to understand the unobserved factors that lead to such negative correlations. Notwithstanding the sign of the correlation terms, the statistically significant correlation parameter highlights the endogeneity of collision type outcomes with injury severity outcomes. Ignoring such endogeneity may lead to poor model fit and biased estimates of variable coefficients. In the current empirical context, the log-likelihood value deteriorated from -15439 (for the joint MNL-ORL model) to -15459 when the collision type was not considered endogenous to the model system. This log-likelihood difference is equivalent to a log-likelihood ratio of 40, which is greater than that the 95% critical chi-square value for 2 degrees of freedom (for the two copula parameters) indicating the statistical superiority of the joint model system.

As can be observed from the table, several of the injury severity model estimates (both thresholds and coefficients) are different across different collision types. This allows the analyst to examine the differential impact of various factors on injury severity by collision type. For example, the threshold values for the head-on collisions are all smaller than the thresholds for other types of collisions. This implies an intuitive result that the probability of higher injury severity categories (fatality etc.) tends to be higher for head-on collisions than all other types of collisions. Further, the snow/ice condition variable has been found to be insignificant for head-on collisions. This implies that the severity of injuries resulting from head-on collisions do not differ by surface conditions, perhaps because of the fact that head-on collisions tend to result in severe accidents anyway due to the high relative speed and impact of the crash. Similar effects can be found in the context of age of vehicle variable that the severity of injuries resulting from head-on collisions does not differ by the age of vehicle. Almost all studies in the literature ignore such differential effects and assume that all factors have the same impact on injury severity irrespective of the type of collision.

Table 4 presents the marginal effects (for the fatal injury category) from the joint model (that considers the endogeneity of collision type) as well as an independent model system (that does not consider the endogeneity of collision type). Several important observations can be made from this table. <u>First</u>, the marginal effects of both the models (independent and joint) show differential impacts of several variables by collision type. This result reiterates the need to examine the differential impact of various factors by collision type. <u>Second</u>, for both model systems, the marginal effects of several variables (gender, age, alcohol/drug use, seat belt use, and environmental and crash characteristics) are smaller in magnitude for head-on collisions than all other types of collisions. This result points to the possibility that the injury severity resulting from head-on collisions tends to be less moderated by various factors. This is especially the case with seat belt use; seat belts are less protective in head-on collisions than in other types of collisions. Since protective measures are less effective in the event of a head-on collision, it is important to reduce the likelihood of a head-on collision in the first place by using control measures such as installation of median barriers. <u>Third</u>, there are non-negligible differences in the marginal effects between the independent and joint models for certain variables. For example, in the context of the medium speed limit variable, the independent model system shows rather small magnitudes of marginal effect when compared to that in the joint model system. On the other hand, the seat belt use variable shows higher marginal effects in the independent model system when compared to that in the joint model system.

5 CONCLUSIONS

This paper employs a recently emerging copula-based methodology to address endogeneity in crash injury severity models. Specifically, two important sources of endogeneity are addressed in the context of two-vehicle collisions: (1) Endogeneity due to the correlations between the injury severities of the two drivers involved in a two-vehicle crash, and (2) Endogeneity of collision type due to the common unobserved factors affecting the collision type and injury severity outcomes. To this end, two sets of joint model systems are formulated and estimated using data on two-vehicle crashes from the Generalized Estimates System (GES) for the year 2007: (1) A copula-based joint ordered logit-ordered logit model of injury severity of the two drivers involved in two-vehicle crashes, and (2) A copula-based joint multinomial logit-ordered logit model of collision type and injury severity outcomes of two-vehicle crashes. To our knowledge, this study constitutes one of the first applications of the copula-based methods to address endogeneity in traffic safety literature.

Model estimation results using the two joint model systems show a statistically significant presence of the two types of endogeneity. Both the model systems provide intuitive results on the impact of various roadway, environmental, vehicle, and driver characteristics on the injury severity of the drivers involved in two-vehicle accidents. Further, the joint model systems perform better than the independent counter parts that do not accommodate the corresponding endogeneity in terms of model fit and policy implications. These results underscore the importance of accommodating endogeneity in crash injury severity modeling, as well as the potential of the copula-based methods in traffic crash modeling and analysis.

Several important findings have surfaced from this analysis. First, drivers under the influence of alcohol/drugs pose a risk of high(fatal) injury not only to themselves but also to other individuals involved in the accident. Such drivers are also found more likely to be involved in head-on collisions. Thus, strict enforcement policies need to be implemented in the context of reducing driving under alcohol/drug influence. Second, the injury severity resulting from head-on collisions tends to be less moderated by other factors (such as seat belt use) than the injuries from other collisions. This result underscores the importance of control measures that reduce the likelihood of head-on collisions.

Given these results and the copula-based method presented in the paper, it is hoped that the issue of endogeneity will be considered with much more importance than earlier in traffic crash analyses. An avenue for future research is to address the extent of (and the importance of) different sources of endogeneity in crash injury severity models. Another research avenue is to investigate the variation in the extent of endogeneity (e.g., the extent of correlations between the injury severity propensities of the two drives involved in a crash) among different crashes.

ACKNOWLEDGEMENTS

The corresponding author is grateful to Naveen Eluru for helpful discussions on copulas. Four anonymous reviewers provided helpful suggestions to improve the paper.

REFERENCES

1. O' Donnell, C.J., and D.H. Connor. Predicting the Severity of Motor Vehicle Accident Injuries Using Models of Ordered Multiple Choice. *Accident Analysis and Prevention*, Vol. 28, No. 6, 1996, pp. 739-753.

2. Dissanayake, S., and J.J. Lu. Factors Influential in Making an Injury Severity Difference to Older Drivers Involved in Fixed Object-Passenger Car Crashes. *Accident Analysis and Prevention*, Vol. 34, No. 5, 2002, pp. 609-618.

3. Washington, S., M. Karlaftis, and F. Mannering. Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall, Boca Raton, 2003.

4. Evans, L. Safety-Belt Effectiveness: The Influence of Crash Severity and Selective Recruitment. *Accident Analysis and Prevention,* Vol. 28, No. 4, 1996, pp. 423-433.

5. Eluru, N., and C.R. Bhat. A joint econometric analysis of seat belt use and crash related injury severity. *Accident Analysis and Prevention*, Vol. 39, No. 5, 2007, pp. 1037–1049.

6. Dee, T.S. Reconsidering the Effects of Seat Belts Laws and their Enforcement Status. *Accident Analysis and Prevention*, Vol. 30, No. 1, 1997, pp. 1-10.

7. Derrig, R.A., M. Segui-Gomez, A. Abtahi, and L.L. Liu. The Effect of Population Safety Belt Usage Rates on Motor Vehicle-Related Fatalities. *Accident Analysis and Prevention,* Vol. 34, No. 1, 2000, pp. 101-110.

8. Cohen, A., and L. Einav. The Effects of Mandatory Seat Belt Laws on Driving Behavior and Traffic Fatalities. *The Review of Economics and Statistics*, Vol. 85, No. 4, 2003, pp. 828-843.

9. de Lapaprent, M. Willingness to use safety belt and levels of injury in car accidents. *Accident Analysis and Prevention*, Vol. 40, No. 3, 2008, pp. 1023-1032.

10. Kim, K., L. Nitz, J. Richardson, and L. Li. Personal and Behavioral Predictors of Automobile Crash and Injury Severity. *Accident Analysis and Prevention,* Vol. 27, No. 4, 1995, pp. 469-481.

11. Ye, X., R.M. Pendyala, F.S. Al-Rukaibi, and K. Konduri. A Joint model of Accident Type and Severity for Two-vehicle Crashes. *Presented at 87th Annual Meeting of the Transportation Research Board,* Washington, D.C., 2008

12. Lee, C., and M. Abdel-Aty. Presence of passengers: does it increase or reduce driver's crash potential? *Accident Analysis and Prevention*, Vol. 40, No. 5, 2008, pp. 1703-1712.

13. Kim, D.G., and S. Washington. The significance of endogeneity problems in crash models: An examination of left-turn lanes in intersection crash models. *Accident Analysis and Prevention,* Vol. 38, No. 6, 2006, pp. 125–134.

14. Hutchinson T.P. A bivariate normal model for intra-accident correlations of driver injury with application to the effect of mass ratio. *Accident Analysis and Prevention,* Vol. 15, No. 3, 1983, pp. 215- 224.

15. Hutchinson, T.P. Statistical Modeling of Injury Severity, with Special Reference to Driver and Front Seat Passenger in Single-Vehicle Crashes. *Accident Analysis and Prevention,* Vol. 18, No. 2, 1986, pp. 157–167.

16. Ouyang, Y., T. Yamamoto, and V.N. Shankar. Modeling the Simultaneity in Injury Causation in Multivehicle Collisions. *Transportation Research Record,* No.1784, 2002, pp. 143–152.

17. Yamamoto, T., and V.N. Shankar. Bivariate ordered-response probit model of driver's and passenger's injury severities in collision with fixed objects. *Accident Analysis and Prevention*, Vol. 36, No. 5, 2004, pp. 869-876.

18. Jones, A.P., and S.H. Jorgensen. The Use of Multilevel Models for the Prediction of Road Accident Outcomes. *Accident Analysis and Prevention,* Vol. 35, No. 1, 2003, pp. 59–69.

19. Lenguerrand, E., J.L. Martin, and B. Laumon. Modeling the hierarchical structure of road crash data: application to severity analysis. *Accident Analysis and Prevention,* Vol. 38, No. 1, 2006, pp. 43-53.

20. Kim, D.G., Y. Lee, S. Washington, and K. Choi. Modeling Crash Outcome Probabilities at Rural Intersections: Application of Hierarchical Binomial Logistic Models. *Accident Analysis and Prevention,* Vol. 39, No. 1, 2007, pp. 1094–1100.

21. Helai, H., C.H. Chor, and M.M. Haque. Severity of Driver Injury and Vehicle Damage in Traffic Crashes at Intersections: A Bayesian Hierarchical Analysis. *Accident Analysis and Prevention,* Vol. 40, No. 1, 2008, pp. 45–54.

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

23. Bhat, C.R., and N. Eluru. A Copula-Based Approach to Accommodate Residential Self-Selection Effects in Travel Behavior Modeling. *Transportation Research Part B*, Vol. 43, No. 7, 2009, pp. 749-765.

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

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

26. Spissu, E., A.R. Pinjari, R.M. Pendyala, C.R. Bhat. A copula-based joint multinomial discrete-continuous model of vehicle type choice and miles of travel. *Transportation*, Vol. 36, No. 4, 2009, pp. 403-422.

27. Kockelman, K.M. and Y.J. Kweon. Driver injury severity: an application of ordered probit models. *Accident Analysis and Prevention,* Vol. 34, No. 3, 2001, pp. 313-321.

28. Gaudry, M., and K. Vernier. Effects of Road Geometry and Surface on Speed and Safety. University of Montreal, Publication AJD-18, 2002.

LIST OF TABLES

Table 1. A Summary of Traffic Safety Literature Addressing Endogeneity

No.	Paper	Data Used	Model Structure Used	Type of Endogeneity	Research Method and Findings
1	Evans (*4*)	National Accident Sampling (NASS) data (1982-91)	Descriptive analysis	Endogeneity of seat belt use	Explored the relationship between the effectiveness of seat belt and crash severity (measured by the change of velocity from before to after crash, inferred from the extent of vehicle damage). Results: Drivers who do not wear seat belts are more likely to be involved in severe crashes, and without considering this effect seat belt effectiveness can be overestimated.
2	Dee (*6*)	Seat belt use data from NHTSA and Center for Disease Control and Prevention (1985-93)	Linear probability models	Endogeneity of seat belt use	Panel data models were used to understand the influence of seat belt laws on fatality rates. Results: Unsafe (or accident prone) drivers are less responsiveness to seat belt laws (*i.e.,* they tend to continue to not wear seatbelts even after the enactment of seat belt laws), which causes an attenuation in the benefits of seat belt laws.
3	Derrig *et al.*(*7*)	Fatality rates and seat belt usage data from Fatality Analysis Reporting System (FARS) (1991-96)	Multivariate regression analysis for per-capita fatality rates per vehicle miles traveled	Endogeneity of seat belt use and state insurance system	Panel data models were used. Instrumental variables were used for the risk-taking incentive of the insurance system by state. Results: Increase in seat belt usage rates in the general population may not lead to reductions in fatality rates, as long as accident prone drivers maintain risky behavior.
4	Cohen and Einav (*8*)	Aggregated data from FARS (1983-97)	Log-linear regression model for per-capita fatality rates in each state in the U.S.	Endogeneity of seat belt use	Presence of mandatory seat-belt laws (by state) was used as an instrument variable to control for the endogeneity of seat-belt usage. Ignoring the endogeneity of seat belt-usage rates resulted in a biased assessment of the effect of seat belt-usage rates on the predicted fatality rates.
5	Eluru and Bhat (*5*)	Generalized Estimates System (GES) data (2003)	Joint random coefficients binary logit - ordered logit model	Endogeneity of seat belt use	Jointly modeled seat belt usage and injury severity. Results indicate the presence of unobserved factors influencing both seat belt use and injury severity. The influence of seat belt use on injury severity was overestimated if the endogeneity of seat belt usage was not considered.
6	de Lapparent (*9*)	French road accident reports (2003)	Bivariate ordered probit model of seat belt use and injury severity	Endogeneity of seat belt use	Jointly modeled seat belt use and injury severity, separately for the drivers, front passengers and rear passengers. Results suggest that while seat belt-use is effective in moderating injury severity, drivers may compensate for some of this safety benefit by taking more risks.
7	Ye *et al.* (*11*)	GES data (2005)	Joint random coefficients multinomial logit - ordered logit model of collision type and injury severity	Endogeneity of collision type variables in two-vehicle crashes	Jointly modeled collision type and injury severity of two-vehicle crash victims. Results suggest that the unobserved factors contributing to head-on collisions are negatively associated with those contributing to severe injuries, where as the unobserved factors contributing to rear end crashes are positively correlated with those contributing to severe injuries
8	Kim *et al.* (*10*)	Crash data from Hawaii CODES project (1990)	Structural Equations Models	Endogeneity of seat belt use and crash type	Structural Equations Model with seat belt use, crash type, and injury severity as endogenous variables. No specific result/discussion was provided on endogeneity.
9	Lee and Abdel-Aty (*12*)	5-year crash records of Interstate-4 freeway in Orlando, Florida (1999–2003)	Bivariate ordered probit models	Endogeneity of passenger characteristics	Jointly modeled passenger characteristics (presence, number, and age of passengers) and crash characteristics (citation, crash type, and injury severity) to capture the endogeneity of passenger characteristics with the crash characteristics (such as injury severity). It was found that drivers display safer driving behavior with the presence of passenger(s), but younger drivers with younger passengers may be more crash prone.
10	Kim and Washington (*13*)	Intersections dataset of 38 counties within the state of Georgia	Joint negative binomial model for angle crashes and logit model for left turn lanes	Endogeneity of left turn lane presence in angle crash occurrence models	Installation of left turn lanes at intersection appears to contribute to crashes when endogeneity is not considered. Recognizing endogeneity results in a negative effect of left turn lanes on occurrence of angle crashes, which is intuitive and concurrent with engineering judgment.
11	Gaudry and Vernier (*28*)	Three different data bases (accident, road and speed) were combined	Logit Model and Linear regression model	Endogeneity of speed	Endogeneity between speed and safety (crash frequency and severity) was considered by three simultaneous equations.

Table 1 (Continued.) A Summary of Traffic Safety Literature Addressing Endogeneity

No.	Paper	Data Used	Model Structure Used	Type of Endogeneity	Research Method and Findings
12	Hutchinson (*14*)	British road accidents data (1969-72)	Bivariate normal distributed model of injury severities of two drivers in a two-vehicle crash	Simultaneity (common unobserved factors influencing injuries of different individuals)	Developed a bivariate ordered probit type model of injury severity of the two drivers involved in a two-vehicle crash. Results indicate that unobserved factors common to both drivers (assumed to be the relative speed in this case) play an important role in determining injury severity levels.
13	Hutchinson (*15*)	British road accidents data (1969-72)	Bivariate normal distributed model of driver's and front passenger's injury severity	Simultaneity (common unobserved factors influencing injuries of different individuals)	Research method same as above. Significant positive correlation was found between the injury severities of the two occupants in a vehicle. The correlation was interpreted a largely due to the speed of the crash.
14	Ouyang *et al.*(*16*)	Washington state accident records database (1990-96)	Simultaneous binary logit model	Simultaneity (common unobserved factors influencing injuries of different individuals)	Jointly modeled the most severe injury in each vehicle for car-truck collisions. Results show significant positive correlation between the two injury severity propensities, and that considering jointness ensures more efficient and less biased estimates.
15	Yamamoto and Shankar (*17*)	Washington state accident records database (1990-96)	Bivariate ordered-response probit model	Simultaneity (common unobserved factors influencing injuries of different individuals)	Jointly modeled the injury severity of the driver and that of the most severely injured passenger in a single vehicle accident. It was found that the error term of the driver's injury severity propensity is positively correlated with that of the most severely injured passenger.
16	Jones and Jørgensen (*18*)	Norwegian road accident police records (1985-96)	Multilevel logistic regression	Common unobserved factors at accident level and accident location level	Three-level (crash victim, crash, and crash location levels) regression models were estimated to disentangle the unobserved factors at the individual, crash, and crash location levels. Results signify the presence of intra-unit correlations in the dataset at both the crash and crash location (municipality, in this case) levels
17	Lenguerrand *et al.* (*19*)	French road crash data (1996-2000)	Multilevel logistic, Generalized Estimating Equations, and simple logistic regression models were compared	Common unobserved factors at vehicle level and accident level	Three-level (crash victim, vehicle, and crash levels) regression models were estimated using multi-level and generalized estimating equations approaches. Results indicate non-negligible correlations at crash level, at the same time indicated a need for large datasets and great care to estimate multilevel models
18	Kim *et al.* (*20*)	Crash data for 91 two-lane rural intersections in the state of Georgia, U.S. (1996–1997)	Multilevel binomial logistic models of crash type	Common unobserved factors at the intersection level	Two-level (crash level and intersection level) binary logistic models of crash type (angle, rear-end, sideswipe, etc.). Results indicate a significant presence of intersection-level unobserved factors affecting the crash type outcomes
19	Helai *et al.* (*21*)	Database of crashes at urban intersections in Singapore	Multi-level logistic regression	Correlation between the individuals involved in same crash	Two-level (individual level and crash level) binary logistic models were estimated. The results show that 28.9% of unexplained variation in severity level results due to the between-crash variance (*i.e.,* crash-level unobserved factors). Significant correlation was found between the accident severities of the individuals involved in the same crash

Table 2: Injury Severity Models for the Two Drivers Involved in Two-Vehicle Crashes: parameter Estimates and Marginal Effects

Variables	Descriptive statistics	Parameter Estimates (t-stats)		Marginal Effects	
		Independent Model (ORL)	Copula-based Joint Model (ORL-ORL)	Independent Model (ORL)	Copula-based Joint Model (ORL-ORL)
Driver Characteristics					
Gender - Female	46.60%	0.55(8.72)	0.53(9.05)	54.26	53.01
Age (> 64 years is base)					
< 25 years	27.00%	-0.58(-5.22)	-0.59(-5.74)	-51.12	-52.43
25 -64 years	63.90%	-0.27(-2.75)	-0.27(-2.88)	-28.71	-28.06
Alcohol or drug use					
Driver	1.90%	0.82(3.65)	0.83(3.82)	124.52	125.61
Driver of partner vehicle	1.90%	0.66(3.10)	0.63(3.07)	91.35	86.42
Use of seat belts	97.40%	-1.65(-9.39)	-1.47(-8.60)	-389.12	-312.89
Roadway Characteristics					
Surface condition (dry/wet is base)					
Snow, ice	3.30%	-0.25(-1.32)	-0.28(-1.35)	-22.11	-24.57
Profile (level is base)					
Grade	18.20%	-0.16(-1.95)	-0.16(-1.83)	-15.18	-15.12
Speed limit (< 26 mph is base)					
Medium (26-65 mph)	85.30%	0.60(6.19)	0.51(5.25)	49.28	43.22
High (> 65 mph)	2.00%	0.41(1.52)	0.32(1.12)	50.00	37.27
Environmental Factors					
Lighting condition (daylight is base)					
Dark	5.30%	0.23(1.75)	0.18(1.21)	25.74	19.56
Land use					
Population > 100000	40.40%	0.31(4.51)	0.29(3.97)	28.59	27.42
Crash Characteristics					
Manner of collision (Rear end & other collision base)					
Head on	6.20%	1.87(14.55)	1.73(12.06)	487.48	417.18
Angle	46.40%	0.58(8.92)	0.51(7.20)	61.09	53.24
Sideswipe collision	7.60%	-0.90(-6.09)	-0.91(-5.73)	-62.01	-62.65
Vehicle role (Striking other vehicle is base)					
Struck by other vehicle	44.40%	0.54(8.46)	0.50(9.64)	55.42	51.55
Struck by and strikes other vehicle	2.30%	1.67(8.04)	1.40(6.77)	409.38	296.19
Vehicle Characteristics					
Body type (Sedan is base)					
Pickup truck	15.80%	-0.48 (-4.9)	-0.49(-5.34)	-40.32	-41.26
Utility vehicle	18.90%	-0.19 (-2.4)	-0.19(-2.51)	-18.10	-17.63
Body type of the Partner vehicle					
Non-sedan (Pickup/Utility vehicle/Minivan)	42.30%	0.18(3.1)	0.20(3.41)	18.79	19.70
Age of vehicle					
> 10 years	26.10%	0.24 (3.4)	0.22(3.52)	25.79	23.46
Threshold Parameters					
Threshold 1	-	1.20(5.26)	1.23(5.53)	-	-
Threshold 2	-	2.49(10.78)	2.51(11.14)	-	-
Threshold 3	-	3.87(16.00)	3.87(16.30)	-	-
Threshold 4	-	6.86(15.32)	6.76(15.65)	-	-
Copula Dependency Parameter (θ)	-	-	1.31(11.04)	-	-
Log Likelihood at convergence	-	-4923.45	-4747.31	-	-

Table 3. Joint MNL-ORL Model of Collision Type and Injury Severity: Parameter Estimates

Variables	Collision Type (MNL) Model Component					Injury Severity (ORL) Model Component				
	Head on	Angle	Rear end	Sideswipe	Others	Head on	Angle	Rear end	Sideswipe	Others
Copula dependency type						Frank	Frank	None	None	None
Copula dependency parameter (θ)	-	-	-	-	-	3.03(5.1)	1.29(3.1)	-	-	-
Constant	-1.59(-13.1)	-0.23(-3.5)	-	-1.07(-11.8)	-3.04(-14.1)	-	-	-	-	-
Threshold (1)	-	-	-	-	-	-1.49(-5.8)	0.37(1.6)	0.53(2.3)	2.05(7.9)	1.52(6.4)
Threshold (2)	-	-	-	-	-	-0.64(-2.5)	1.52(6.4)	1.82(7.5)	3.19(10.1)	2.38(6.5)
Threshold (3)	-	-	-	-	-	0.37(1.6)	2.92(11.6)	2.92(11.6)	4.44(9.9)	4.44(9.9)
Threshold (4)	-	-	-	-	-	2.38(6.5)	6.01(10.1)	6.76(6.4)	6.76(6.4)	6.76(6.4)
Driver Characteristics										
Gender										
Female	-	-	-	-	-	0.49(8.5)	0.49(8.5)	0.49(8.5)	0.49(8.5)	0.49(8.5)
Age (> 64 years is base)										
< 25 years	-	-	-	-	-	-0.51(-4.9)	-0.51(-4.9)	-0.51(-4.9)	-0.51(-4.9)	-0.51(-4.9)
25 -64 years	-	-	-	-	-	-0.24(-2.6)	-0.24(-2.6)	-0.24(-2.6)	-0.24(-2.6)	-0.24(-2.6)
Alcohol or drug use										
Driver	0.75(2.8)	-	-	-	-	0.75(3.6)	0.75(3.6)	0.75(3.6)	0.75(3.6)	0.75(3.6)
Driver of partner vehicle	-	-	-	-	-	0.62(3.2)	0.62(3.2)	0.62(3.2)	0.62(3.2)	0.62(3.2)
Use of seat belts	-	-	-	-	-	-1.50(-8.9)	-1.50(-8.9)	-1.50(-8.9)	-1.50(-8.9)	-1.50(-8.9)
Roadway Characteristics										
Surface condition (dry / wet is base)										
Snow, ice	0.77(2.2)	-	0.58(2.9)	0.60(2.1)	-	-	-0.49(-1.7)	-0.28(-1.1)	-	-
Profile (level is base)										
Grade	-	-	-	-	-	-0.15(-2.1)	-0.15(-2.1)	-0.15(-2.1)	-0.15(-2.1)	-0.15(-2.1)
Alignment (straight is base)										
Curve	0.35(2.2)	-0.93(-8.9)	-	-	-	-	-	-	-	-
Speed limit (< 26 mph is base)										
Medium (26-65 mph)	-	-	-	-	-	0.94(3.9)	0.54(4.9)	0.44(3.3)	0.44(3.3)	0.44(3.3)
High (> 65 mph)	-	-	-	-	-	0.36(1.5)	0.36(1.5)	0.36(1.5)	0.36(1.5)	0.36(1.5)
Absence of median	1.20(7.6)	0.56(10.8)	-	-	-	-	-	-	-	-
Number of lanes - 3or more	-	0.15(3.2)	-	0.24(3.4)	-	-	-	-	-	-

Table 3 (Continued.) Joint MNL-ORL Model of Collision Type and Injury Severity: Parameter Estimates

Variables	Collision Type (MNL) Model Component					Injury Severity (ORL) Model Component				
	Head on	Angle	Rear end	Sideswipe	Others	Head on	Angle	Rear end	Sideswipe	Others
Environmental Factors										
Lighting condition										
Daylight	-	-	-	-	-	-	-	-	-	-
Dark	0.54(4.2)	0.12(2.1)	-	0.19(2.2)	0.49(2.0)	0.24(1.4)	0.24(1.4)	-	0.24(1.4)	0.24(1.4)
Land use										
Population > 100000	-	0.09(1.9)	-	0.10(1.3)	0.65(2.5)	0.27(4.3)	0.27(4.3)	0.27(4.3)	0.27(4.3)	0.27(4.3)
Crash Characteristics										
Vehicle role (Striking is base)										
Struck	-	-	-	-	-	0.50(8.6)	0.50(8.6)	0.50(8.6)	0.50(8.6)	0.50(8.6)
Struck by and strikes	-	-	-	-	-	1.63(8.2)	1.63(8.2)	1.63(8.2)	1.63(8.2)	1.63(8.2)
Accident at intersection	-	1.08(15.9)	0.58(8.2)	-	-	-	-	-	-	-
Traffic Control Device										
Stop sign	-	0.98(12.6)	-	-	-	-	-	-	-	-
Stop sign not at intersection	-	1.39(5.3)	-	-	-	-	-	-	-	-
Yield sign	-	1.06(2.9)	2.12(6.6)	-	-	-	-	-	-	-
Traffic control signal	-	-	0.31(5.2)	-	-	-	-	-	-	-
Vehicle Characteristics										
Body type (Sedan is base)										
Pickup truck	-	-	-	-	-	-0.84(-2.5)	-0.57(-4.0)	-0.33(-2.7)	-	-
Utility vehicle	-	-	-	-	-	-	-0.20(-1.7)	-0.32(-3.1)	-	-
Body type of the Partner vehicle										
Non-sedan (Pickup/Utility vehicle)	-	-	-	-	-	0.26(3.3)	0.26(3.3)	-	-	-
Age of vehicle										
> 10 years	-	-	-	-	-	-	0.24(3.5)	0.24(3.5)	-	-

Table 4. Marginal Effects for Fatal Injury Severity from Independent and Joint Collision Type and Injury Severity Model Systems

Variables	Marginal Effects from Independent Model System					Marginal Effects from Joint Model System				
	Head on	Angle	Rear end	Side- swipe	Others	Head on	Angle	Rear end	Side- swipe	Others
Driver Characteristics										
Gender										
Female	52.36	52.88	53.69	53.16	60.62	45.82	48.13	48.94	48.51	54.52
Age (> 64 years is base)										
< 25 years	-50.80	-51.54	-50.40	-50.79	-48.57	-43.75	-46.12	-45.23	-45.55	-43.78
25 -64 years	-27.53	-28.42	-29.06	-28.92	-29.51	-22.78	-24.71	-25.23	-25.09	-25.55
Alcohol or drug use										
Driver	109.36	119.32	119.51	120.06	112.00	93.16	108.71	108.97	109.36	102.66
Driver of partner vehicle	89.22	95.57	95.47	95.22	93.58	73.38	83.96	84.01	83.77	82.41
Use of seat belts	-327.04	-379.94	-390.00	-387.73	-406.32	-249.96	-323.42	-330.99	-329.71	-342.19
Roadway Characteristics										
Surface condition (dry / wet is base)										
Snow, ice	-	-38.44	-33.03	-	-	-	-39.44	-24.52	-	-
Profile (level is base)										
Grade	-16.15	-16.44	-16.61	-16.48	-16.74	-13.40	-14.26	-14.39	-14.29	-14.49
Speed limit (< 26 mph is base)										
Medium (26-65 mph)	28.51	59.04	33.41	33.23	33.46	69.11	46.04	37.42	37.22	37.57
High (> 65 mph)	34.53	36.00	35.86	35.79	35.24	38.49	42.65	42.48	42.37	41.63
Environmental Factors										
Lighting condition										
Dark	33.26	35.21	-	35.10	33.52	24.26	26.84	-	26.78	25.87
Land use										
Population > 100000	27.81	28.17	28.48	28.16	27.34	23.35	24.73	24.95	24.70	24.08
Crash Characteristics										
Vehicle role (Striking is base)										
Struck by other vehicle	68.60	56.26	54.56	58.14	63.12	56.48	50.37	49.14	51.79	55.84
Struck by and strikes other vehicle	392.26	436.94	460.55	410.93	375.64	299.22	387.40	406.13	367.14	336.82
Vehicle Characteristics										
Body type (Sedan is base)										
Pickup truck	-72.26	-44.10	-38.38	-	-	-60.11	-46.12	-29.16	-	-
Utility vehicle	-	-14..57	-35.28	-	-	-	-18.72	-28.59	-	-
Body type of the Partner vehicle										
Non-sedan (Pickup , utility)	28.02	29.22	-	-	-	24.26	26.35	-	-	-
Age of vehicle										
> 10 years	-	28.37	29.02	-	-	-	24.34	24.82	-	-