Accident Analysis and Prevention 45 (2012) 110–119

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

A multivariate tobit analysis of highway accident-injury-severity rates

Panagiotis Ch. Anastasopoulos[a,*], Venky N. Shankar[b,1], John E. Haddock[c,2], Fred L. Mannering[d,3]

[a] School of Civil Engineering, Purdue University, Indiana Local Technical Assistance Program, 550 Stadium Mall Drive, West Lafayette, IN 47907-2051, United States
[b] Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, PA 16802, United States
[c] Department of Civil Engineering, Purdue University, and Indiana Local Technical Assistance Program, 550 Stadium Mall Drive, West Lafayette, IN 47907-2051, United States
[d] Department of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907-2051, United States

ARTICLE INFO

Article history:
Received 23 August 2011
Received in revised form 14 October 2011
Accepted 9 November 2011

Keywords:
Accident-injury-severity rates
Censoring
Contemporaneous error correlation
Multivariate tobit regression
Roadway geometrics
Interstate highways

ABSTRACT

Relatively recent research has illustrated the potential that tobit regression has in studying factors that affect vehicle accident rates (accidents per distance traveled) on specific roadway segments. Tobit regression has been used because accident rates on specific roadway segments are continuous data that are left-censored at zero (they are censored because accidents may not be observed on all roadway segments during the period over which data are collected). This censoring may arise from a number of sources, one of which being the possibility that less severe crashes may be under-reported and thus may be less likely to appear in crash databases. Traditional tobit-regression analyses have dealt with the overall accident rate (all crashes regardless of injury severity), so the issue of censoring by the severity of crashes has not been addressed. However, a tobit-regression approach that considers accident rates by injury-severity level, such as the rate of no-injury, possible injury and injury accidents per distance traveled (as opposed to all accidents regardless of injury-severity), can potentially provide new insights, and address the possibility that censoring may vary by crash-injury severity. Using five-year data from highways in Washington State, this paper estimates a multivariate tobit model of accident-injury-severity rates that addresses the possibility of differential censoring across injury-severity levels, while also accounting for the possible contemporaneous error correlation resulting from commonly shared unobserved characteristics across roadway segments. The empirical results show that the multivariate tobit model outperforms its univariate counterpart, is practically equivalent to the multivariate negative binomial model, and has the potential to provide a fuller understanding of the factors determining accident-injury-severity rates on specific roadway segments.

Published by Elsevier Ltd.

1. Introduction

The preponderance of past research has studied the occurrence of accidents by considering their frequency and applying count-data modeling techniques to study factors that affect the frequency of accidents over some time period on specific roadway segments. This body of literature has applied a wide variety of modeling approaches such as Poisson and negative binomial models, Poisson-lognormal models, zero-inflated count models, Conway–Maxwell–Poisson models, negative multinomial models, Gamma models, generalized estimating equation models, generalized additive models, random effects and random parameters count models, and finite mixture and Markov switching models (for a complete review of this literature see Lord and Mannering, 2010).

While traditional accident-frequency approaches have undeniably improved our understanding of factors affecting accident occurrence, some recent research has suggested the tobit regression as an alternative (Anastasopoulos et al., 2008, 2012). The tobit-regression approach considers accident rates (such as the number of accidents per vehicle-miles traveled) on roadway segments as opposed to accident frequencies. This results in data that are continuous (instead of the discrete count data in the traditional frequency approaches) and in data that are left-censored at zero because accidents may not be reported on some roadway segments during the time period over which data are collected. This censoring may occur for a number of reasons ranging from the simple possibility that no accidents occurred on the roadway segment over the study period, to the possibility that accidents not involving injury may not be reported if their property damage does not exceed a specified threshold (this threshold is open to the interpretation of the officer at the accident scene and thresholds may vary from one jurisdiction to the next).

* Corresponding author. Tel.: +1 765 494 7760.
E-mail addresses: panast@purdue.edu (P.Ch. Anastasopoulos), shankarv@engr.psu.edu (V.N. Shankar), jhaddock@purdue.edu (J.E. Haddock), flm@purdue.edu (F.L. Mannering).
[1] Tel.: +1 814 865 9434.
[2] Tel.: +1 765 496 3996.
[3] Tel.: +1 765 496 7913.

0001-4575/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.aap.2011.11.006

In the traditional tobit framework, which models overall accident rates irrespective of their injury severity, the equivalent effect of these censoring sources is assumed to be the same across all accident-injury-severity levels – which can be problematic if no-injury accidents, for example, are less likely to be reported. A solution to this would be to model the accident rates by injury-severity level. However, if accident rates by injury-severity level are modeled independently, significant estimation error could be introduced because unobserved effects at the roadway-segment level are likely to be shared across severities. This general problem of shared unobserved effects[4] has been previously encountered in traditional accident-frequency modeling where, instead of modeling the total number of accidents, separate count-data models are estimated for each accident-injury type. This has led to the application of bivariate models, where two accident types are considered (Subrahmaniam and Subrahmaniam, 1973; Maher and Summersgill, 1996; N'Guessan et al., 2006; Geedipally and Lord, 2010; N'Guessan, 2010), and multivariate models, where three or more accident types are considered (Winkelmann, 2008; Bijleveld, 2005; Ma and Kockelman, 2006; Song et al., 2006; Park and Lord, 2007; Ma et al., 2008; El-Basyouny and Sayed, 2009; Park et al., 2010). The intent of this paper is to address both issues – censoring masking, and contemporaneous error correlation of the latent variables due to commonly shared unobserved characteristics – simultaneously, by demonstrating the multivariate tobit model as an approach of dealing with correlation among injury-severity models when the data are continuous and censored (accident rates on roadway segments) as opposed to the more traditional count-frequency data.

2. Methodology

Tobit regression (Tobin, 1958) is used with left-censored (censored at a low threshold) or right-censored (censored at a high threshold) dependent variables. For the accident rate of specific injury-severity levels (defined by the degree of injury of the most severely injured occupant in the accident), the data can be left-censored with a clustering at zero (zero accidents of a specified injury-severity per 100-million vehicle miles traveled) because, as discussed earlier, accidents of a specific injury severity may not be observed simply because none have occurred or due to the non-availability of data (accidents that do not involve injury, typically are only reported if the property value damage exceeds a pre-specified value threshold). In tobit regression, the equivalent effect is assumed to be the same, clustered-at-zero observations.

To address the possibility of multiple tobit regressions by injury-severity level (no-injury crashes per 100-million vehicle miles traveled, possible-injury per 100-million vehicle miles traveled, and injury crashes per 100-million vehicle miles traveled), past research that has dealt with estimation techniques to model multiple tobit equations with contemporaneous (cross-equation) error correlation (Huang et al., 1987; Huang, 1999; Trivedi and Zimmer, 2005) can be used to develop a multivariate tobit model for accident rates on specific roadway segments[5] considering accidents whose most severely injured vehicle occupant was either not injured, possibly injured, or injured.[6] Using a left-censored limit of zero, the multivariate tobit model with three dependent variables is expressed as:

$$\begin{aligned} Y_{ik}^* &= \mathbf{X}_{ik}'\boldsymbol{\beta}_k + \varepsilon_{ik},\ i = 1, 2, \ldots, N, \quad k = 1, 2, 3 \\ Y_{ik} &= Y_{ik}^* \qquad\qquad \text{if}\ \ Y_{ik}^* > 0 \\ &= 0 \qquad\qquad\ \ \text{if}\ \ Y_{ik}^* \leq 0, \end{aligned} \qquad , \tag{1}$$

where N is the number of observations, Y_{ik} is the dependent variable for the kth accident-injury-severity rate (1–3 for no-injury, possible injury, and injury accidents per 100-million vehicle miles traveled in roadway segment i, respectively) for the ith segment, $\mathbf{X}_{ik}'$ is a vector of independent variables (pavement, traffic, weather and roadway segment characteristics), $\boldsymbol{\beta}_k$ is a vector of estimable parameters, and ε_{ik} are multivariate normally and independently distributed error terms with zero mean, variance σ^2, correlation ρ, and covariance matrix:

$$\Sigma_{\varepsilon_k} = \begin{pmatrix} \sigma_{\varepsilon_1}^2 & \rho_{\varepsilon_2\varepsilon_1}\sigma_{\varepsilon_2}\sigma_{\varepsilon_1} & \rho_{\varepsilon_3\varepsilon_1}\sigma_{\varepsilon_3}\sigma_{\varepsilon_1} \\ \rho_{\varepsilon_1\varepsilon_2}\sigma_{\varepsilon_1}\sigma_{\varepsilon_2} & \sigma_{\varepsilon_2}^2 & \rho_{\varepsilon_3\varepsilon_2}\sigma_{\varepsilon_3}\sigma_{\varepsilon_2} \\ \rho_{\varepsilon_1\varepsilon_2}\sigma_{\varepsilon_1}\sigma_{\varepsilon_3} & \rho_{\varepsilon_2\varepsilon_3}\sigma_{\varepsilon_2}\sigma_{\varepsilon_3} & \sigma_{\varepsilon_3}^2 \end{pmatrix}. \tag{2}$$

And given these error terms, the density function of Y_{ik} is (Trivedi and Zimmer, 2005):

$$fk(Y_{ik}|\mathbf{X}_{ik}\boldsymbol{\beta}_k') = \prod_{Y_{ik}=0}\left[1 - \Phi\left(\frac{\mathbf{X}_{ik}'\boldsymbol{\beta}_k}{\sigma_k}\right)\right]\prod_{Y_{ik}>0}\phi\left(\frac{Y_{ik} - \mathbf{X}_{ik}'\boldsymbol{\beta}_k}{\sigma_k}\right), \tag{3}$$

where Φ is the multivariate normal distribution function and ϕ is the multivariate normal density function. Eq. (1) shows that there is an implicit, stochastic index (latent variable) equal to Y_{ik}^* which is observed only when positive.

The multivariate (trivariate) tobit distribution is:

$$F(Y_1, Y_2, Y_3) = C[F_1(Y_{i1}|\mathbf{X}_{i1}'\boldsymbol{\beta}_1), F_2(Y_{i2}|\mathbf{X}_{i2}'\boldsymbol{\beta}_2), F_3(Y_{i3}|\mathbf{X}_{i3}'\boldsymbol{\beta}_3); \theta], \tag{4}$$

and the corresponding log-likelihood function for the multivariate (trivariate) tobit model is:

$$\begin{aligned} &L_N[(Y_1|\mathbf{X}_1; \boldsymbol{\beta}_1), (Y_2|\mathbf{X}_2; \boldsymbol{\beta}_2), (Y_3|\mathbf{X}_3; \boldsymbol{\beta}_3); \theta] \\ &\quad = \sum_{i=1}^{N}\sum_{k=1}^{3}\ln f_{ik}\left(Y_{ik}|\mathbf{X}_{ik}; \boldsymbol{\beta}_k\right) + \sum_{i=1}^{N} C_{123}\left[F_1\left(Y_{i1}|\mathbf{X}_{i1}; \boldsymbol{\beta}_{i1}\right),\right. \\ &\quad\quad \left. F_2\left(Y_{i2}|\mathbf{X}_{i2}; \boldsymbol{\beta}_{i2}\right), F_3\left(Y_{i3}|\mathbf{X}_{i3}; \boldsymbol{\beta}_{i3}\right); \theta\right], \end{aligned} \tag{5}$$

where $C_{123}(\cdot)$ is the cross partial derivative for the copula (the function linking marginal variables into the multivariate distribution; see Trivedi and Zimmer (2005) and Prokhorov and Schmidt (2009) for specifics on the copula estimation), θ the dependence

[4] Such unobserved effects may be capturing driver-specific information (age, gender, marital status, socioeconomic status, risk taking, driving experience, driving behavior, driving adjustment in situational responses, etc.), or information on vehicle characteristics, such as vehicle type, safety features (airbags, anti-lock brakes, etc.), and horsepower (see Janssen, 1994; Dee, 1998; Winston et al., 2006). Such information is typically available only after an accident has occurred and thus cannot be used to predict accident rates.

[5] Note that modeling accident rates using tobit regression is an appropriate solution, in order to avoid biased and inconsistent parameter estimates that would otherwise be the result of the ordinary least squares (OLS) estimation which does not account for the censoring in the data. The same limitations exist when modeling censored data as a system of equations (such as seemingly unrelated regression equations, two-stage least squares, or three-stage least squares), which implicitly assume that the zero observations are generated by the same process that generates the positive observations (see Washington et al., 2011). The univariate and multivariate tobit models by definition assume different generation processes for the zero and non-zero observations (see Eq. (1)), which makes them suitable for the analysis and modeling of left-censored, continuous data, as the accident rates.

[6] Evident injury rates include injury severity levels classified as evident injury, disabling injury and fatal injury. The disabling and fatal injury rates individually resulted in a sparse column of positive rates, while disabling and fatal injuries combined did not significantly improve the column sparseness of non-zero values. It is acknowledged that this is a limiting assumption in that correlations are constrained to be the same within evident, disabling and fatal injury categories due to the composite definition.