Analytic Methods in Accident Research 16 (2017) 23–34

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Simultaneous estimation of discrete outcome and continuous dependent variable equations: A bivariate random effects modeling approach with unrestricted instruments

Md Tawfiq Sarwar [a], Grigorios Fountas [b], Panagiotis Ch. Anastasopoulos [c,*]

[a] National Research Council, Federal Highway Administration, Turner-Fairbank Highway Research Center, 6300 Georgetown Pike, T-210 McLean, VA 22101, United States

[b] Department of Civil, Structural and Environmental Engineering, Engineering Statistics and Econometrics Application Research Laboratory, University at Buffalo, The State University of New York, 204B Ketter Hall, Buffalo, NY 14260, United States

[c] Department of Civil, Structural and Environmental Engineering, Institute for Sustainable Transportation and Logistics, Engineering Statistics and Econometrics Application Research Lab, University at Buffalo, The State University of New York, 241 Ketter Hall, Buffalo, New York 14260, United States

ARTICLE INFO

Article history:
Received 17 February 2017
Received in revised form 9 May 2017
Accepted 9 May 2017

Keywords:
System of equations
Bivariate model
Continuous dependent variable
Discrete outcome variable (binary)
Endogeneity
Cross-equation correlation
Panel effects
Pedestrian crosswalks

ABSTRACT

This paper proposes a novel methodology to simultaneously model discrete outcome (binary) and continuous dependent variables. The proposed modeling framework addresses unobserved heterogeneity by accounting for both panel effects, and for contemporaneous (cross-equation) error correlation between the two dependent variables; while, variable endogeneity is addressed through the use of unrestricted – equation specific – instruments. To illustrate the applicability of the bivariate modeling framework, SHRP2 Naturalistic Driving Study (NDS) data are used to empirically investigate the driving behavior preceding pedestrian crosswalks, in terms of brake application (binary dependent variable, binary probit specified) and speed change (continuous dependent variable, linear regression specified), simultaneously. The bivariate model is counter-imposed against its univariate binary probit and linear regression counterparts. The results of the comparative assessment demonstrate the statistical superiority of the proposed bivariate modeling approach – in terms of explanatory power, statistical fit, and forecasting accuracy – and its potential in modeling multivariate mixed dependent variables.

Published by Elsevier Ltd.

1. Introduction

Over the recent years, technological and methodological advancements have paved the way towards multivariate modeling settings, characterized by large datasets and cutting-edge but computationally manageable estimation techniques. Such multivariate modeling schemes (i.e., modeling frameworks that jointly estimate multiple but homogeneous dependent variables) address significant limitations of the traditional univariate modeling approaches, such as: interrelationships among the dependent variables (Washington et al., 2011; Mannering and Bhat, 2014); commonly shared unobserved effects among the dependent variables, which can result in contemporaneous correlation of disturbances (Washington et al., 2011; Sarwar and Anastasopoulos, 2016; Mothafer et al., 2016); spatial or temporal dependency, which can result in subsequent correlation effects (Mannering and Bhat, 2014); and presence of endogeneity – arising from the interdependency of the

* Corresponding author.
E-mail addresses: mdtawfiq@buffalo.edu (M.T. Sarwar), gfountas@buffalo.edu (G. Fountas), panastas@buffalo.edu (P.Ch. Anastasopoulos).

http://dx.doi.org/10.1016/j.amar.2017.05.002
2213-6657/Published by Elsevier Ltd.

dependent variables – which can result in the presence of correlation between the endogenous dependent variables and the equation error terms (Kim and Washington, 2006; Washington et al., 2011; Sarwar and Anastasopoulos, 2016). In words, the multivariate framework allows for modeling multiple dependent variables simultaneously, while at the same time controls for potential model specification and estimation issues. This is important, since not accounting for model misspecifications such as endogeneity or cross-equation error correlation, can result in inconsistent parameter estimates, inaccurate predictions, and erroneous inferences (Washington et al., 2011).

In this context, a growing amount of research has been devoted on the estimation of simultaneous equations models (SEM). Specifically, full information maximum likelihood (FIML) techniques (Savolainen and Mannering, 2007; Washington et al., 2011), three-stage least squares (3SLS) models (Anastasopoulos et al., 2010; Bhargava et al., 2010; Sarwar and Anastasopoulos, 2016), and seemingly unrelated regression equations (SURE) models (Mannering, 2007; Anastasopoulos et al., 2012a; Anastasopoulos and Mannering, 2016; Sarwar and Anastasopoulos, 2017) constitute common multivariate approaches for continuous dependent variables. For truncated or censored continuous data, as in the accident rates analysis, multivariate tobit models have been primarily used (Anastasopoulos et al., 2012c; Anastasopoulos, 2016; Sarwar and Anastasopoulos, 2017). For count data dependent variables, multivariate count data models have been widely implemented in the safety related literature, especially under the prism of modeling accident frequencies by injury-severity level or accident type (Ye et al., 2009, 2013; Anastasopoulos et al., 2012c; Dong et al., 2014; Barua et al., 2015, 2016; Anastasopoulos, 2016; Cheng et al., 2017). For ordinal scale data, multivariate ordered probit approaches have been adopted (Yamamoto and Shankar, 2004; Bhat and Srinivasan, 2005; Eluru et al., 2010; Anastasopoulos et al., 2012b; Abay et al., 2013; Chiou et al., 2013; Russo et al., 2014); whereas, for binary discrete outcome dependent variables, multivariate probit models have been employed (Sarwar et al., 2017a).

A common limitation of the multivariate approaches found in the literature, is associated with the homogeneous (uniform) nature of the dependent variables, under which the latter are modeled simultaneously with statistical models of the same functional form. This restriction can be relaxed through the simultaneous estimation of mixed dependent variables with statistical models of different functional forms.[1] Among the possible combinations of simultaneous model estimation of mixed dependent variables, the joint investigation of continuous and discrete outcome variables has considerable appeal.[2] However, in the absence of a multivariate distribution that inter-connects such mixed dependent variables under different functional forms, model estimation becomes a challenging and complex task (Bhat, 2015).

To that end, this paper provides a new – to the authors' knowledge – modeling framework that simultaneously accommodates both discrete outcome and continuous dependent variables (e.g., discrete binary outcome and continuous dependent variables, modeled jointly with binary probit and linear regression, respectively). The proposed bivariate model accounts for commonly encountered model misspecification issues in highly dimensional data, such as endogeneity, cross-equation error correlation, and panel effects.[3] To demonstrate the applicability of the proposed framework in the context of pedestrian safety, driving behavior data from the SHRP2 Naturalistic Driving Study (NDS) are used and analyzed, while consideration is given to two dependent variables related to brake pedal application (discrete binary outcome dependent variable) and vehicle speed (continuous dependent variable). To provide additional insights regarding the efficiency of the proposed approach, the bivariate model estimation results are compared to its univariate linear regression and binary probit counterparts.

2. Methodology

The proposed bivariate modeling framework is formulated by developing an integrated estimation procedure, on the basis of popular estimation techniques that deal with continuous and discrete binary outcome equations with contemporaneous (cross-equation) error correlation and possible endogeneity (Maddala, 1983; Greene, 2012). Specifically, the bivariate model is expressed as:

$$y_1^* = \gamma_1 y_2^* + \boldsymbol{\beta}_1' \mathbf{X}_1 + u_1, \quad y_1 = y_1^*, \tag{1}$$

$$y_2^* = \gamma_2 y_1^* + \boldsymbol{\beta}_2' \mathbf{X}_2 + u_2, \quad y_2 = 1\ (y_2^* > 0). \tag{2}$$

where, y_1^* and y_2^* denote latent variables, with y_1^* being directly observed as a continuous variable y_1, and y_2^* being observed as a discrete binary outcome variable y_2; $\mathbf{X}_1$ and $\mathbf{X}_2$ are vectors of explanatory variables; $\boldsymbol{\beta}_1'$ and $\boldsymbol{\beta}_2'$ are vectors of estimable parameters; γ are estimable scalars of the endogenous variables; and u's are independently distributed disturbance terms, with variances ${\sigma_1}^2$ and ${\sigma_2}^2$, respectively, and cross-equation error correlation ρ_{12}. Because y_2 is a binary variable, it is

[1] Past research has simultaneously modeled discrete and ordered outcome variables (Bhat and Guo, 2007; Eluru and Bhat, 2007; Abay et al., 2013), or has employed Copula-based model formulations in an effort to account for the effect of dependency between the equations' disturbance terms (Spissu et al., 2009; Bhat and Eluru, 2009).

[2] Simultaneous modeling of discrete and continuous dependent variables can allow for identification of factors affecting the likelihood and amount (or extent): (a) of the same variable (e.g., the likelihood of snowfall, and the amount of snow on the ground); or (b) of different variables (e.g., the likelihood of snowfall, and the temperature). This is important especially in empirical or exploratory studies, because the investigation of one problem dimension (for example, either amount or likelihood) in a univariate framework is unlikely to warrant unbiased and efficient parameter estimates, robust inferences, and accurate forecasts, as compared to multivariate modeling.

[3] It should be noted that, in the proposed bivariate modeling framework, endogeneity is addressed through an unrestricted instruments approach, under which the instruments are estimated with different sets of exogenous variables (even though, the same exogenous variables can also be used).