Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

A count data model with endogenous covariates: Formulation and application to roadway crash frequency at intersections

Chandra R. Bhat [a,b,*], Kathryn Born [c,1], Raghuprasad Sidharthan [d,2], Prerna C. Bhat [e,3]

[a] *The University of Texas at Austin, Department of Civil, Architectural and Environmental Engineering, 301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, USA*
[b] *King Abdulaziz University, Jeddah 21589, Saudi Arabia*
[c] *The University of Texas at Austin, Department of Civil, Architectural and Environmental Engineering, 301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, USA*
[d] *Parsons Brinckerhoff, 999 3rd Ave, Suite 3200, Seattle, WA 98104, USA*
[e] *Harvard University, 1350 Massachusetts Avenue, Cambridge, MA 02138, USA*

ARTICLE INFO

Article history:
Received 16 August 2013
Received in revised form
15 October 2013
Accepted 16 October 2013
Available online 15 November 2013

Keywords:
Count data
Treatment-outcome models
Accident analysis
Generalized ordered response
Flashing light control

ABSTRACT

This paper proposes an estimation approach for count data models with endogenous covariates. The maximum approximate composite marginal likelihood inference approach is used to estimate model parameters. The modeling framework is applied to predict crash frequency at urban intersections in Irving, Texas. The sample is drawn from the Texas Department of Transportation (TxDOT) crash incident files for the year 2008. The results highlight the importance of accommodating endogeneity effects in count models. In addition, the results reveal the increased propensity for crashes at intersections with flashing lights, intersections with crest approaches, and intersections that are on frontage roads.

1. Introduction

This paper develops an estimation approach for count data models with endogenous covariates, where the endogenous covariates are based on a multinomial probit discrete outcome model. The proposed formulation model constitutes a specific version of the generalized Roy model that is referred to as the treatment effects model (see Heckman and Vytlacil, 2005 and Bhat and Eluru, 2009). In the empirical context studied in the paper, the type of control at an intersection constitutes the treatment. The type of control is represented in five categories: (1) no traffic control (including intersections with no control and intersections with some minimal form of control such as turn marks and marked lanes), (2) yield sign control on one more approaches with no other form of control, (3) stop sign control on one or more approaches, (4) flashing light control (one or more approaches having a flashing red or yellow light), and (5) regular signal light control. The count outcome in the empirical context is the number of crashes at urban intersections. In this case, the type of traffic control may itself be determined by the frequency of crashes, as, in fact, is explicitly noted in the Manual on Uniform Traffic Control Devices or MUTCD (FHWA, 2009). For instance, the total entering volume of traffic on the approach roadways to an intersection may directly impact both the type of control as well as the frequency of crashes, creating an "endogeneity" of the type of traffic

[*] Corresponding author at: The University of Texas at Austin, Department of Civil, Architectural and Environmental Engineering, 301 E. Dean Keeton St. Stop C1761, Austin, Texas 78712, United States. Tel.: +512 471 4535; fax: +512 475 8744.
E-mail addresses: bhat@mail.utexas.edu (C.R. Bhat), born2@utexas.edu (K. Born), srprasad@utexas.edu (R. Sidharthan), prernabhat@college.harvard.edu (P.C. Bhat).
[1] Tel.: +512 471 4535; fax: +512 475 8744.
[2] Tel.: +206 382 5289; fax: +206 382 5222.
[3] Tel.: +512 289 0221.

2213-6657/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.amar.2013.10.001

control in crash frequency analysis. But if the entering volume were an observed variable, then this type of "endogeneity" is easily accommodated by including the entering volume as an explanatory variable, along with traffic control type, in the modeling of crash frequency. More generally, if the determination of the control types at intersections were random conditional on observed characteristics, a traditional count model for crash frequency would suffice. However, many unobserved factors may affect both control type and crash frequency, rendering the random conditional (on observed characteristics) assumption untenable. For instance, at intersections with an unobserved terrestrial/topographic feature that limits sight distance, flashing lights may be installed instead of a stop sign. That same unobserved feature may be responsible for a lower or a higher frequency of crashes (one can argue that motorists are more careful when they encounter some unobserved topographic feature, resulting in a lower frequency of crashes; alternatively, it could also be that the unobserved feature results in a higher frequency of crashes). If one of these two situations exists, but is ignored, it would generate an inconsistent, spurious, and biased effect of flashing lights on intersection crash frequency. Of course, there are many other application contexts where our proposed model formulation should be useful, including the type of insurance plan ("treatment") and the number of doctor visits (outcome), residential location type ("treatment") and the number of cancer incidents (outcome), the intensity of lighting at a train station and the number of crimes at the station, and one of many other contexts where count models are used to model the outcome decision. However, in the development of the model formulation in this paper, as in the empirical analysis context of the paper, the focus will be on traffic control type and the frequency of crashes at urban intersections.

Methodologically speaking, our parametric multinomial discrete-count model uses a general multinomial probit (MNP) specification for the treatment and ties this MNP model with a count model. In particular, we use Castro et al., 2012 recasting of a univariate count model as a restricted version of a univariate generalized ordered-response probit (GORP) system. In addition to providing substantial flexibility to accommodate high or low probability masses for specific count outcomes, the latent variable-based count specification of the GORP system provides a convenient mechanism to tie the count outcome with the MNP treatment model. In this regard, our proposed model (which we will refer to as the Count model with endogenous multinomial probit selection, or the CEMPS model) has some similarity with Bhat (1998) and Munkin and Trivedi's (2008) ordered probit model with endogenous selection, but with four important differences. First, the outcome variables in the earlier models were ordinal variables, while the outcome variable in our CEMPS model is a true count variable that can take on any non-negative integer value. Second, the earlier models did not allow random response variations (or unobserved heterogeneity) in the sensitivity to exogenous factors in both the selection (or treatment) component as well as the outcome component. On the other hand, it is now well established that ignoring such response variations when present will lead to inconsistent and biased parameters estimates in both multinomial discrete response models as well as count models (see Chamberlain, 1980, Bhat, 1998). For instance, variations in the effect of entering volume on the type of control installed at an intersection and on crash frequency may result from the complex interactions between unobserved intersection characteristics and motorist learning/adaptation behavior in response to different levels of traffic volume. Accommodating such unobserved heterogeneity effects is not simply an esoteric econometric effort, but can have very real implications for accurately assessing the overall effects of variables on the outcome of interest (for example, to design countermeasures to reduce crash frequency in our empirical context; see Anastasopoulos and Mannering, 2009, Castro et al., 2013, and Mannering and Bhat, 2014). Third, we allow unobserved heterogeneity in the treatment effects themselves rather than *a priori* positing fixed treatment effects. For example, even after controlling for endogeneity effects, the "true" effect of control type on crash frequency itself may vary across intersections due to such unobserved intersection geometric features as curb radii and approach configuration. Fourth, unlike Bhat (1998), we use an MNP-based treatment model rather than a very restrictive multinomial logit treatment model, and use a simple frequentist inference approach rather than Munkin and Trivedi's relatively cumbersome Bayesian estimation approach. The frequentist approach is based on an analytic (as opposed to a simulation) approximation of the multivariate normal cumulative distribution (MVNCD) function that appears in the full likelihood function of the proposed model. Bhat (2011) discusses this analytic approach, which is based on earlier works by Solow (1960) and Joe (1996). The approach involves only univariate and bivariate cumulative normal distribution function evaluations in the likelihood function.

In summary, and to our knowledge, this is the first formulation and application of a flexible count outcome model with a multinomial probit selection model, which also accommodates unobserved heterogeneity effects.

2. Model formulation

2.1. The selection (treatment) MNP model

In the usual random discrete response model formulation, write the unobserved continuous random latent variable influencing the probability that intersection q is controlled by traffic control type i as follows:

$$U_{qi} = \boldsymbol{\beta}_q^{'} \mathbf{x}_{qi} + \xi_{qi} \tag{1}$$

where $\mathbf{x}_{qi}$ is a $(D \times 1)$-column vector of exogenous attributes (including a dummy variable for each control type alternative except a base control type), $\boldsymbol{\beta}_q$ is an individual-specific $(D \times 1)$-column vector of corresponding coefficients that varies across intersections based on unobserved intersection attributes, and ξ_{qi} captures the idiosyncratic (unobserved) intersection characteristics that impact the latent propensity of control type i being installed at intersection q (in the rest of this paper, we will refer to U_{qi} as the propensity of control type i being installed at intersection q). We assume that the error terms ξ_{qi} are multivariate normally distributed across control types i for a given intersection q: $\boldsymbol{\xi}_q = (\xi_{q1}, \xi_{q2}, \ldots, \xi_{qI})' \sim MVN_I(\mathbf{0}_I, \boldsymbol{\Lambda})$, where $MVN_I(\mathbf{0}_I, \boldsymbol{\Lambda})$ indicates an I-variate normal distribution with a mean vector of zeros denoted by $\mathbf{0}_I$ and a covariance matrix $\boldsymbol{\Lambda}$. Such a specification captures the possibility that,