Analytic Methods in Accident Research 1 (2014) 72–85

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models

Fan Ye [a,*], Dominique Lord [b,1]

[a] Texas A&M Transportation Institute, Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135, USA
[b] Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, USA

ARTICLE INFO

Article history:
Received 22 February 2013
Received in revised form
3 March 2013
Accepted 3 March 2013
Available online 22 May 2013

Keywords:
Sample size
Crash severity model
Multinomial logit model
Ordered probit model
Mixed logit model

ABSTRACT

There have been many studies that have documented the application of crash severity models to explore the relationship between accident severity and its contributing factors. Although a large amount of work has been done on different types of models, no research has been conducted about quantifying the sample size requirements for crash severity modeling. Similar to count data models, small data sets could significantly influence model performance. The objective of this study is therefore to examine the effects of sample size on the three most commonly used crash severity models: multinomial logit, ordered probit and mixed logit models. The study objective is accomplished via a Monte-Carlo approach using simulated and observed crash data. The results of this study are consistent with prior expectations in that small sample sizes significantly affect the development of crash severity models, no matter which type is used. Furthermore, among the three models, the mixed logit model requires the largest sample size, while the ordered probit model requires the lowest sample size. The sample size requirement for the multinomial logit model is located between these two models.

Published by Elsevier Ltd.

1. Introduction

Discrete response models in traffic safety (often referred to as crash severity models), such as logit and probit models, are usually used to explore the relationship between accident severity and its contributing factors such as driver characteristics, vehicle characteristics, roadway conditions, and road-environment factors. A review of these types of models that have been used for crash severity analyses shows that they can be generally classified as either nominal or ordinal (see Savolainen et al. (2011) for a thorough review). Among the nominal models, the three most common ones are multinomial logit models, nested logit models, and mixed logit models. The ordinal models, on the other hand, can also be classified into three groups: ordered logit models, ordered probit models, and ordered mixed logit models. There are other types of crash severity models, but they are not as popular or used in practice. The curious reader is referred to Savolainen et al. (2011) for an extensive list of available models for analyzing crash severity. Overall, based on the existing literature, the multinomial logit models and ordered probit models have been found to be the most prominent types of models used for traffic crash severity analysis (see Table 1 in

[*] Corresponding author. Tel.: +1 979 845 7415.
E-mail addresses: f-ye@ttimail.tamu.edu (F. Ye), d-lord@tamu.edu (D. Lord).
[1] Tel.: +1 979 458 3949.

2213-6657/$ - see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.amar.2013.03.001

Savolainen et al. (2011)). Meanwhile, the mixed logit model is a promising model that has recently been used widely in many different areas.

Few research studies have been conducted on directly comparing different crash severity models, though each model type has its own unique benefits and limitations. So far, there is no consensus on which model is the best, as the selection of the model is often governed by the availability and characteristics of the data (Savolainen et al., 2011). Some researchers prefer choosing nominal models over ordinal models because of the restriction placed on how variables affect ordered discrete outcome probabilities; that is using the same coefficient for a variable among different crash severities. Others still prefer ordinal models due to its simplicity and overall performance when less detailed data are available (Washington et al., 2011). From the few researchers who directly compared crash severity models, Abdel-Aty (2003) recommended the ordered probit model over the multinomial logit models and nested logit models, while Haleem and Abdel-Aty (2010) reported that the aggregate binary probit model (a special case of an ordered probit model by aggregating the five crash severity levels into two) offered superior performances compared to the ordered probit and nested logit models in terms of goodness-of-fit.

Similar to count data models (Lord, 2006), crash severity models can be heavily influenced by the size of the sample from which they are estimated. As discussed in previous research (Lord and Bonneson, 2005; Lord and Mannering, 2010), crash data are often characterized by a small number of observations. This attribute is credited to the large costs of assembling crash and other related data. Although it is anticipated that the size of the sample will influence the performance of crash severity models, nobody has so far quantified how the sample size affects the most commonly used crash severity models and consequently provide guidelines on the data size requirements. A few have proposed such guidelines, but only for crash-frequency models (Lord, 2006; Lord and Miranda-Moreno, 2008; Park et al., 2010). In addition, crash severity models are usually estimated using the maximum-likelihood estimator (MLE), which is a consistent estimator (ensuring that standard errors of parameter estimates become smaller as sample size becomes larger), but not necessarily an efficient estimator (i.e., for a given sample size the parameter estimate may not have the lowest possible standard error), thus estimation results can be problematic in small samples (Washington et al., 2011).

As stated above, there is a need to examine how sample size can influence the development of commonly used crash severity models. Providing this information could help transportation safety analysts in their decision to use one model over another given the size and characteristics of the data. The objective of this study is therefore to examine the effects of sample size on the three most commonly used crash severity models: the multinomial logit, ordered probit and mixed logit models. The objective is accomplished using a Monte-Carlo analysis based on simulated and observed data. The sample sizes analyzed varied from 100 to 10,000 observations.

2. Methodological background

This section describes the three crash severity models: the multinomial logit, ordered probit, and mixed logit models. The multinomial logit model is derived under the assumption that the unobserved factors are uncorrelated over the alternatives or outcomes, also known as the independence from irrelevant alternatives (IIA) assumption (Train, 2003). This assumption is the most notable limitation of the multinomial logit model since it is very likely that the unobserved factors are shared by some outcomes. Despite this limitation, the IIA assumption makes the multinomial logit model very convenient to use which also explains its popularity.

In the general case of a multinomial logit model of crash injury severity outcomes, the propensity of crash i towards severity category k is represented by severity propensity function,T_{ki}, as shown in Eq. (1) (Kim et al., 2008).

$$T_{ki} = \alpha_k + \boldsymbol{\beta}_k \mathbf{X}_{ki} + \varepsilon_{ki} \tag{1}$$

where, α_k is a constant parameter for crash severity category k; $\boldsymbol{\beta}_k$ is a vector of the estimable parameters for crash severity category k; $k=1,\ldots, K$ ($K=5$ in the paper) representing all the five severity levels: no-injury (NI), possible injury (PI), non-incapacitating injury (NII), incapacitating injury (II), and fatal (F); $\mathbf{X}_{ki}$ represents a vector of explanatory variables affecting the crash severity for i at severity category k (geometric variables, environmental conditions, driver characteristics, etc.); ε_{ki} is a random error term following the Type I generalized extreme value (i.e., Gumbel) distribution; $i=1,\ldots,n$ where n is the total number of crash events included in the model.

Eq. (2) shows how to calculate the probability for each crash severity category. Let $P_i(k)$ be the probability of accident i ending in crash severity category k, such that

$$P_i(k) = \frac{\exp(\alpha_\mathrm{k} + \boldsymbol{\beta}_k \mathbf{X}_{ki})}{\sum_{\forall k} \exp(\alpha_\mathrm{k} + \boldsymbol{\beta}_k \mathbf{X}_{ki})} \tag{2}$$

The ordered probit model uses a latent variable z, as shown in Eq. (3) to determine crash-severity outcomes.

$$z = \boldsymbol{\beta}\mathbf{X} + \varepsilon \tag{3}$$

where $\mathbf{X}$ is a vector of explanatory variables for the individual crash; $\boldsymbol{\beta}$ is a vector of the coefficients for the explanatory variables; and ε is a random error term following standard normal distribution.