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

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

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.

Using Eq. (3), the value of the dependent variable y is determined by

$$y = \begin{cases} 1, & \text{if } z \le \gamma_1 \\ k, & \text{if } \gamma_{k-1} < z \le \gamma_k \\ K, & \text{if } z > \gamma_{K-1} \end{cases} \tag{4}$$

where, $\gamma = \{\gamma_1, \ldots, \gamma_k, \ldots, \gamma_{K-1}\}$ are the threshold values for all crash severity categories, corresponding to integer ordering; $k = 1, \ldots, K$ ($K = 5$ in the paper), representing all the five severity levels: 1 = no-injury (NI), 2 = possible injury (PI), 3 = non-incapacitating injury (NII), 4 = incapacitating injury (II), and 5 = fatal (F); K is the highest ordered crash severity category.

Given the value of $\mathbf{X}$, the probability that the crash severity of an individual crash belongs to each category is

$$\begin{cases} P(y = 1) = \Phi(-\boldsymbol{\beta}\mathbf{X}) \\ P(y = k) = \Phi(\gamma_{k-1} - \boldsymbol{\beta}\mathbf{X}) - \Phi(\gamma_{k-2} - \boldsymbol{\beta}\mathbf{X}) \\ P(y = K) = 1 - \Phi(\gamma_{K-1} - \boldsymbol{\beta}\mathbf{X}) \end{cases} \tag{5}$$

where, $\Phi(\cdot)$ stands for the cumulative probability function of the standard normal distribution.

As stated by Eluru et al. (2008), the standard ordered response models (including the ordered probit model) have a limitation in that the threshold values are fixed across observations, which could lead to inconsistent model estimation. Therefore, these authors introduced a new type of model known as the mixed generalized ordered response logit model for analyzing crash data. The mixed generalized ordered response logit model can generalize the standard ordered response models by allowing the flexibility of the effects of covariates on the threshold value for each ordinal category. However, given the complexity of the model and the fact that it has only been used once, the mixed generalized ordered response logit model was not examined in this study. Furthermore, Abdel-Aty et al. (2011) used a new model, the multilevel ordered logistic model, to study the effects of fog/smoke on crashes. The multilevel ordered logistic model is an extension of ordinary ordered logit model, accounting for the cross-segment heterogeneities by including a random effect component in the thresholds. Using the same method, ordered probit models could be extended into multilevel ordered probit ones in future research, which is beyond the scope of this paper.

The mixed logit model has attracted considerable attention by traffic safety researchers because of its flexibility in model definition and it has become popular due to the improvement in computer power and the development of simulation techniques which are necessary for model estimations (Milton et al., 2008). Mixed logit probabilities are the integrals of standard logit probabilities over a density of parameters (i.e., it is a weighted average of the logit formula evaluated at different value of parameters ($\boldsymbol{\beta}$), with the weights given by the density $f(\boldsymbol{\beta})$).

The mixed logit model shares the same structure of severity propensity function, T_{ki}, utilized for the multinomial logit model, as shown in Eq. (1). Therefore, Eq. (6) shows the calculation of the probability of each crash severity category for the mixed logit model.

Let $P_i(k)$ be the probability of accident i ending in crash severity category k, such that

$$P_i(k) = \int \frac{\exp\left[\alpha_k + \boldsymbol{\beta}_k \mathbf{X}_{ki}\right]}{\sum_{\forall k} \exp(\alpha_k + \boldsymbol{\beta}_k \mathbf{X}_{ki})} f(\boldsymbol{\beta}|\boldsymbol{\theta}) d\boldsymbol{\beta} \tag{6}$$

where, $f(\boldsymbol{\beta}|\boldsymbol{\theta})$ is the density function of $\boldsymbol{\beta}$ with $\boldsymbol{\theta}$ referring to a vector of parameters of the density function (mean and variance).

3. Data

The primary data sources utilized in this study included four years (from 1998 to 2001) of traffic crash records provided by the Texas Department of Public Safety and the Texas Department of Transportation general road inventory. This research investigated the probability of crash severities of single-vehicle traffic accidents involving fixed objects that occurred on rural two-way highways (excluding those occurring at intersections). There were a total of 26,175 usable records in the database which contained a variety of information including conditions of weather, roadway, driver and vehicle as well as crash severities reported at the time of the accidents. For this dataset, these categories had 11,844 (45.3%), 5270 (20.1%), 5807 (22.2%), 2449 (9.4%), and 805 (3.1%) observations for severity no-injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal, respectively. There were 27 independent variables used in the empirical analysis which are summarized in Table 1.

4. Model estimation results

Using the above data, three models (the multinomial logit, ordered probit and mixed logit models) were developed, estimating the probabilities of the five crash severity levels conditional on an accident having occurred. For model estimation, LIMDEP 9.0 was used (Greene, 2007). The estimation results for each model are listed in Table 2. In addition, they are briefly explained as follows and readers are referred to original document for additional information of three models estimation (Ye, 2011).

Table 1
Summary statistics for the variables included in the models.

Variable type	Description	Mean	St.d
Road condition			
Log(ADT)	Log of average daily traffic	7.597	0.999
Shoulder width in feet	Shoulder width varied between 0 and 20 ft	4.865	3.264
Lane width in feet	Lane width varied between 8 ft and 16 ft	11.341	1.251
Speed limit in mi/h	Maximum speed limit varied between 30 mi/h and 75 mi/h	58.330	6.935
Curve and level indicator	1=curve, level; 0=otherwise	0.373	0.484
Curve and grade indicator	1=curve, grade; 0=otherwise	0.002	0.048
Curve and hill indicator	1=curve, hill; 0=otherwise	0.002	0.047
Accident information			
Night indicator	1=night;0=day	0.495	0.500
Dark with no light indicator	1=dark with no light; 0=otherwise	0.424	0.494
Dark with light indicator	1=dark with light; 0=otherwise	0.033	0.177
Rain indicator	1=rain; 0=otherwise	0.806	0.395
Snow indicator	1=snow; 0=otherwise	0.005	0.068
Fog indicator	1=fog; 0=otherwise	0.023	0.149
Surface condition indicator	0=good surface(dry); 1=otherwise	0.267	0.442
Driver information			
Vehicle type indicator	1=truck; 0=otherwise	0.474	0.499
Driver gender indicator	1=female; 0=male	0.340	0.474
Driver's age	in years	32.743	15.245
Driver defect indicator	1=defect (including physical and mental defect); 0=otherwise	0.176	0.381
Restraining device use indicator	1=no restraining device used; 0=otherwise	0.120	0.325
Fatigue indicator	1=fatigued or asleep; 0=otherwise	0.151	0.358
Airbag deploy indicator	1=air bag deployed; 0=otherwise	0.179	0.384
Seat belt use indicator	1=seat belt used; 0=otherwise	0.649	0.477
Fixed-object type information			
Hit pole indictor	1=hit pole; 0=otherwise	0.113	0.317
Hit tree indictor	1=hit tree; 0=otherwise	0.224	0.417
Hit fence indictor	1=hit fence; 0=otherwise	0.261	0.439
Hit bridge indictor	1=hit bridge; 0=otherwise	0.052	0.222
Hit barrier indictor	1=hit barrier; 0=otherwise	0.058	0.233

4.1. Analysis for the multinomial logit model

In the procedure of estimating the multinomial logit model, all 27 explanatory variables mentioned in Table 1 were tested for inclusion, but only 10 variables were retained, as shown in Table 2. The criteria used for variables inclusion were data availability, engineering judgment, and significance level (0.05 is used in this study). For the five crash severity levels, fatal was used as the baseline outcome. Initially, coefficients of a variable in the severity propensity function T_{ki} were specified to be different across all four severity categories (except for fatal, as a baseline outcome). If no significant difference at a 0.05 significance level was observed among the coefficients in two of the severity propensity functions, then they were set to be equal. Likelihood ratio tests were used to test whether the coefficients of a variable in the four severity propensity functions were significantly different from each other. In addition, the Small-Hsiao IIA test (Washington et al., 2011) was conducted. Based on the test, the multinomial logit model structure cannot be refuted and IIA assumption among the five crash severities could not be rejected at the 0.10 significance level for the dataset.

4.2. Analysis for the ordered probit model

For the ordered probit model estimation, all the 27 variables mentioned above were initially included in the model, and the backward selection was used for the selection process so that only those significant at the 0.05 significance level were included in the model. For the final result, more significant variables were kept (18 variables) than the multinomial logit and mixed logit models (10 variables). The signs and values for the estimated coefficients of variables from all three models were deemed reasonable. Sensitivity analyses and direct elasticities that support the interpretation of these variables were performed for each model, but are not presented here due to space limitations (see Ye, 2011 for more details).

4.3. Analysis for the mixed logit model

The mixed logit model allows for the randomness of the parameters of a variable, and thus in developing the model, we first assumed all parameters included in the model which were random. The popular distributions (normal, uniform and lognormal

distribution) were tested for the random parameters, so numerous combinations of these distributions were evaluated by modifying the parameter assumptions. Then, the t-test was used to examine their estimated standard deviations for exploring the randomness of each parameter: if their standard deviation was not found statistically different from zero at the 0.05 significance level, they were restricted to be fixed instead of random. The simulation-based maximum likelihood method was used for parameter estimation, with 200 Halton draws (Milton et al., 2008). The final result of the mixed logit model estimation, as shown in Table 2, was based on engineering judgment and goodness-of-fit measurement.

4.4. Model results comparison

Based on the output of the three models, it is found that the mixed logit model is more interpretive than the multinomial logit model, since the former includes the randomness associated with parameters of some variables in propensity functions, rather than being fixed for each variable by allowing both a mean and standard deviation. That is to say, depending on the parameter distribution, the parameter effects for the mixed logit model can vary across individual crash, ranging from positive to negative and of varying magnitudes (Milton, 2006). This results in the prediction of a mean value and standard deviation for the probability of each severity level rather than a single point probability. Meanwhile, though accounting for the ordinal information of crash severities, the ordered probit model still does not have the same good interpretive power as the multinomial logit and mixed logit models. The ordered probit model restricts the effects of explanatory variables on ordered discrete outcome probabilities by using the identical coefficient for an explanatory variable across different crash severities. It causes the variable either to increase the probability of highest severity (fatal in the study) and decrease the probability of lowest severity (no-injury in the study), or to decrease the probability of highest severity and increase the probability of lowest severity. However, it does not allow the probabilities of both of the highest and lowest severity increase or decrease. This may not be realistic because it is possible that some explanatory variables can create an increase in the probability for some outcome predictions but decrease the probability for other outcome predictions. For instance, inclement weather could lead to an increase in the probability for both highest severities (fatal, incapacitating injury) and lowest severity (no-injury), but reduce the probability of the other severities (possible injury, non-incapacitating injury). In addition, it is not clear what effects a positive or negative variable parameter has on the probabilities of the "interior" severity levels: incapacitating injury, non-incapacitating injury, and possible injury (Washington et al., 2011).

In terms of the goodness-of-fit among three models, the ordered probit model includes more significant variables (18 variables) which results in a slightly higher adjusted rho-squared value (Adjusted $\rho^2 = 0.208$) than those of the multinomial logit and mixed logit model (Adjusted $\rho^2 = 0.194$). Since the multinomial logit model is a nested model of the mixed logit model, we can further compare their goodness-of-fit using a likelihood ratio test, even though both of them have the same adjusted rho-squared value. From the multinomial logit model estimation results in Table 2, the log-likelihood at convergence is −33,926.2 with 27 estimated parameters (degrees of freedom, including four estimated constant variables), and the log-likelihood at convergence for the mixed logit model estimation is −33,917.3 with 30 estimated parameters (three more randomness in the variables than the multinomial logit model). Therefore, the likelihood ratio statistic is $2 \times (-33{,}917.3 - (-33{,}926.2)) = 8.9$ with 3 degrees of freedom, which is larger than the χ^2 table value of 7.81 for the 0.05 level of significance. This indicates that the mixed logit model is statistically better than the multinomial logit model in terms of goodness-of-fit at the 0.05 significance level.

5. Model comparisons by sample size

For this part of the analysis, we used simulated data as well as the four-year accident records described above. Recall that this dataset includes 26,175 single-vehicle accidents involving fixed objects on rural two-way highways. Intuitively, small sample size in crash severity models can lead to erratic results, which limit their ability to estimate the true parameters and result in an inaccurate prediction of the probabilities for each severity outcome. In order to find the difference in sample size requirements for the three models discussed above, a Monte Carlo simulation was used to examine the potential bias associated with different sample sizes for each model type.

5.1. Analysis based on simulated data

By repeating the sampling to produce estimators more clustered around the true values (designed values for the simulated data), the Monte-Carlo simulation is an ideal way to verify the sample size effects on three models since we create the data with the knowledge of true values of estimators and true response functions. Thus, the bias can all be attained by comparing the model estimation with the true values of estimators for different sample sizes.

5.1.1. Simulation design

All the variables included in a crash severity model are observation-related rather than outcome-related, which means that the variables keep the same values no matter what accident severity the target observed crash is (Khorashadi et al., 2005). In other words, the variables in the propensity functions for each severity category for an observed crash are identical though their parameters which describe the effects of crash characteristics might differ across each severity. Thus, the covariate in the propensity functions generated in the simulation should be kept the same for all severities in each observation.

Table 2
Estimation Results of the multinomial logit, ordered probit and mixed logit models based on the observed crash data.

Variable[a]	Multinomial logit				Ordered probit	Mixed logit			
	NI[b]	PI	NII	II		NI	PI	NII	II
Constant	4.489 (12.0)[c]	4.166 (11.1)	3.816 (10.2)	3.213 (9.3)	0.249 (2.8)	4.430 (11.5)	4.155 (10.8)	3.764 (9.8)	3.235 (9.2)
Road condition									
log(ADT)	0.153 (7.5)	0.074 (3.7)	0.074 (3.7)		−0.049 (−6.9)	0.167 (7.7)	0.079 (3.7)	0.079 (3.7)	
Speed limit	−0.02 (−3.8)	−0.02 (−3.8)	−0.02 (−3.8)	−0.02 (−3.8)	0.002 (2.47)	−0.02 (−3.8)	−0.02 (−3.8)	−0.02 (−3.8)	−0.02 (−3.8)
Curve and level indicator					0.062 (4.18)				
Accident information									
Night indicator		−0.229 (−6.8)	−0.153 (−5.2)	−0.153 (−5.2)	−0.124 (−4.4)		−0.238 (−6.9)	−0.183 (−5.2)	−0.183 (−5.2)
Dark with light indicator	0.152 (2.1)					0.166 (2.0)			
Rain indicator	−0.93 (−7.2)	−0.819 (−6.2)	−0.523 (−4.0)	−0.39 (−2.8)		−0.939 (−7.1)	−0.81 (−6.1)	−0.997 (−4.3)	−0.397 (−2.8)
Std.dev. of distribution								1.568 (4.0)	
Snow indicator	0.473 (2.4)					0.466 (2.4)			
Fog indicator					0.106 (2.3)				
Surface condition indicator					−0.259 (−15.7)				
Drive information									
Vehicle type indicator					0.0561 (3.8)				
Driver gender indicator					0.132 (8.6)				
Driver defect indicator	−1.26 (−10.0)	−0.28 (−3.3)	−0.28 (−3.3)	−0.28 (−3.3)	0.398 (9.4)	−1.359 (−9.8)	−0.24 (−2.7)	−0.24 (−2.7)	−0.24 (−2.7)
Restraining device used indicator	−2.53 (−30.8)	−1.99 (−23.2)	−1.40 (−17.5)	−0.83 (−9.77)	0.802 (21.8)	−3.406 (−7.6)	−2.0 (−23.2)	−1.25 (−11.9)	−0.834 (−9.8)
Std.dev. of distribution						2.22 (3.0)			
Fatigue indicator	0.465 (4.6)	−0.258 (−5.2)			−0.173 (−3.8)	0.507 (4.4)	−0.33 (−5.7)		
Airbag deploy indicator					0.447 (12.6)				
Seat belt use indicator					−0.128 (−3.9)				
Fixed-object type information									
Hit pole indictor					−0.076 (−3.2)				
Hit tree indicator	−1.05 (−13.2)	−0.83 (−10.1)	−0.612 (−7.6)	−0.36 (−4.2)	0.188 (10.1)	−1.14 (−11.8)	−0.86 (−10.3)	−0.561 (−6.3)	−0.378 (−4.4)
Std.dev. of distribution						0.939 (2.4)			
Hit fence indictor					−0.16 (−8.8)				
Hit barrier indictor					−0.09 (−2.9)				
Threshold parameters									
y_1					0.561 (86.2)				
y_2					1.393 (139.6)				
y_3					2.186 (133.2)				
Log-likelihood at zero	−42,127.0				−42,127.0	−42,127.0			
Log-likelihood at convergence	−33,926.2				−33,328.9	−33,917.3			
Adjusted ρ^2	0.194				0.208	0.194			

[a] See Table 1 for full variable description.
[b] NI: no-injury; PI: possible injury; NII: non-incapacitating injury; II: incapacitating injury.
[c] Values in parentheses are the *t*-ratio of each estimated parameter.

Table 3
True parameter values used in the simulation for the three models.

Model parameter		True values		
		Multinomial logit	Ordered probit	Mixed logit
Constant parameter[a]	PI[b] constant	1.5	2.4	1.5
	NII constant	1	1.5	1
	II constant	0.5	0.8	0.5
	F constant	0	0	0
Variable parameter	PI variable	1	1	1
	NII variable	1		1
	II variable	1		1
	F variable	1		N(1,1)
Sample size (N)		100, 250, 500, 1000, 1500, 2000, 5000, 10,000		

[a] Constant parameter for the ordered probit model is represented by γ_1–γ_4, which are the threshold variables for each outcome in the ordered probit model.
[b] NI: no-injury; PI: possible injury; NII: non-incapacitating injury; II: incapacitating injury.

Since the crash data have five severity categories, the number of parameters to investigate is very large. For simplification, one covariate randomly generated from the standard normal distribution was introduced for all three models. In addition, five outcomes (denoted as injury levels 1 to 5, from no-injury to fatal) will be used to replicate the five severity categories.

The three datasets for each model were generated. For the multinomial logit model, the variable parameters were kept the same with a value equal to 1 for each outcome, i.e., $\beta_k = 1$. Constant parameters α_k were 1.5, 1, 0.5, 0 for injury levels 2–5 (level 1, no-injury, was the baseline outcome with $\alpha_1 = \beta_1 = 0$). The independent variable x for each level was drawn from a normal distribution with mean equal to −2 and a variance equal to 1. The error term for each level was drawn independently from a Type I extreme value distribution by obtaining draws from the uniform random distribution and applying the following transformation $-\ln[-\ln(u)]$, where u was a random number drawn from the uniform distribution between 0 and 1. Thus, they gave the following proportions 44.1%, 25.4%, 15.4%, 9.4% and 5.7% for injury levels 1–5 correspondingly which represented the proportions observed in the data (five crash severities from no-injury to fatal).

For the ordered probit model, the variable parameter β was equal to 1 for each level, x was drawn from a normal distribution with a mean equal to 2.2 and a variance equal to 1, and threshold variables γ_k were 2.4, 1.5, 0.8, 0 for levels 2–5 (for keeping the population ratios of each level as close as those for the multinomial logit model). The error term was standard normally distributed for each outcome. Thus, they gave the following proportions 44.3%, 24.6%, 15.0%, 10.1%, and 6.0% for injury levels 1–5 (from no-injury to fatal), correspondingly.

For the mixed logit model, the steps for generating the dataset were very similar to those used in generating the dataset for the multinomial logit model. The only difference is that the independent variable was assumed to have a random component in the variable parameter for injury level 1 (no-injury), which followed a normal distribution (mean = 1, variance = 1). The population proportions for each outcome were 39.3%, 23.6%, 14.3%, 8.7% and 14.1% for injury levels 1–5 (from no-injury to fatal), correspondingly. What can be noticed is that the proportions of each level for the mixed logit model are not as close as those for the multinomial logit and ordered probit models. This can be attributed to the existing randomness associated with the mixed logit model. The randomness causes more variability of the data and makes the proportions harder to be controlled.

Table 3 summarizes the true values assumed for three models. The parameter values chosen for three models were based on the assumption the results would not be affected much by different values of the parameters.

Datasets of each model were repeatedly drawn 100 times for each sample size according to the designed true parameter values of the model. The sample sizes were designed as 100, 250, 500, 1000, 1500, 2000, 5000, and finally 10,000.

5.1.2. Simulation results

Turning first to the simulation results for the multinomial logit model, based on quintiles from the empirical sampling distribution of the parameter estimators, 95% confidence intervals were calculated for each estimated parameter. The graphs in Fig. 1 show the relationship between 95% confidence intervals for the four estimated constant parameters and the parameters associated with the independent variables for the sample sizes described above. In each graph, the Y-axis is the parameter estimate, and the X-axis is the sample size. For each sample size, there are two estimates of the parameter, one for lower-bound and the other for upper-bound of the 95% confidence intervals. Thus, the interval encloses a 95% probability of the real value of each parameter.

From Fig. 1, it can be noticed that for each parameter, the range for the 95% confidence interval becomes narrower as the sample size gets larger, though no direct inverse proportional relationship has been found between the 95% confidence interval and sample size. In addition, as the sample size reaches 2000, the 95% confidence interval gets smaller and stays stable around the true value for each parameter. In order to take a closer look at the simulation results, the relationship between the mean value of each parameter and sample size was extracted and is illustrated in Fig. 2. This figure shows that sample sizes less than

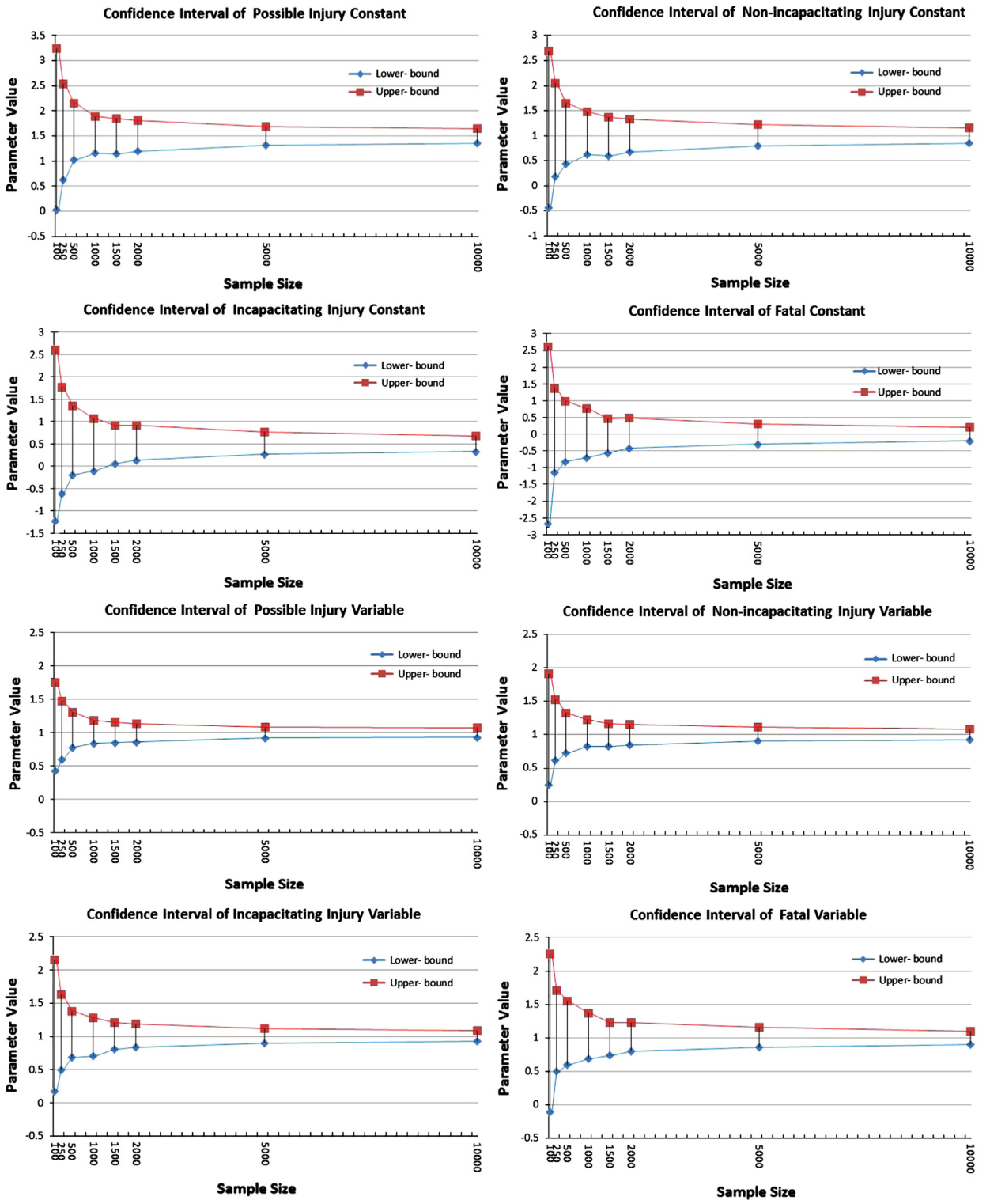

Fig. 1. Confidence intervals of the parameters by sample size for the multinomial logit model.

2000 are somewhat erratic in the abilities to find the true parameters. Furthermore, the estimated mean value for all the variables appears to be biased for all four coefficients. At this point, the factors influencing the bias are unknown and additional work is needed to determine what causes this bias. The mean value becomes stable for a sample size greater than 2000, which is about the same value when the 95% confidence interval becomes much smaller, as seen in Fig. 1.

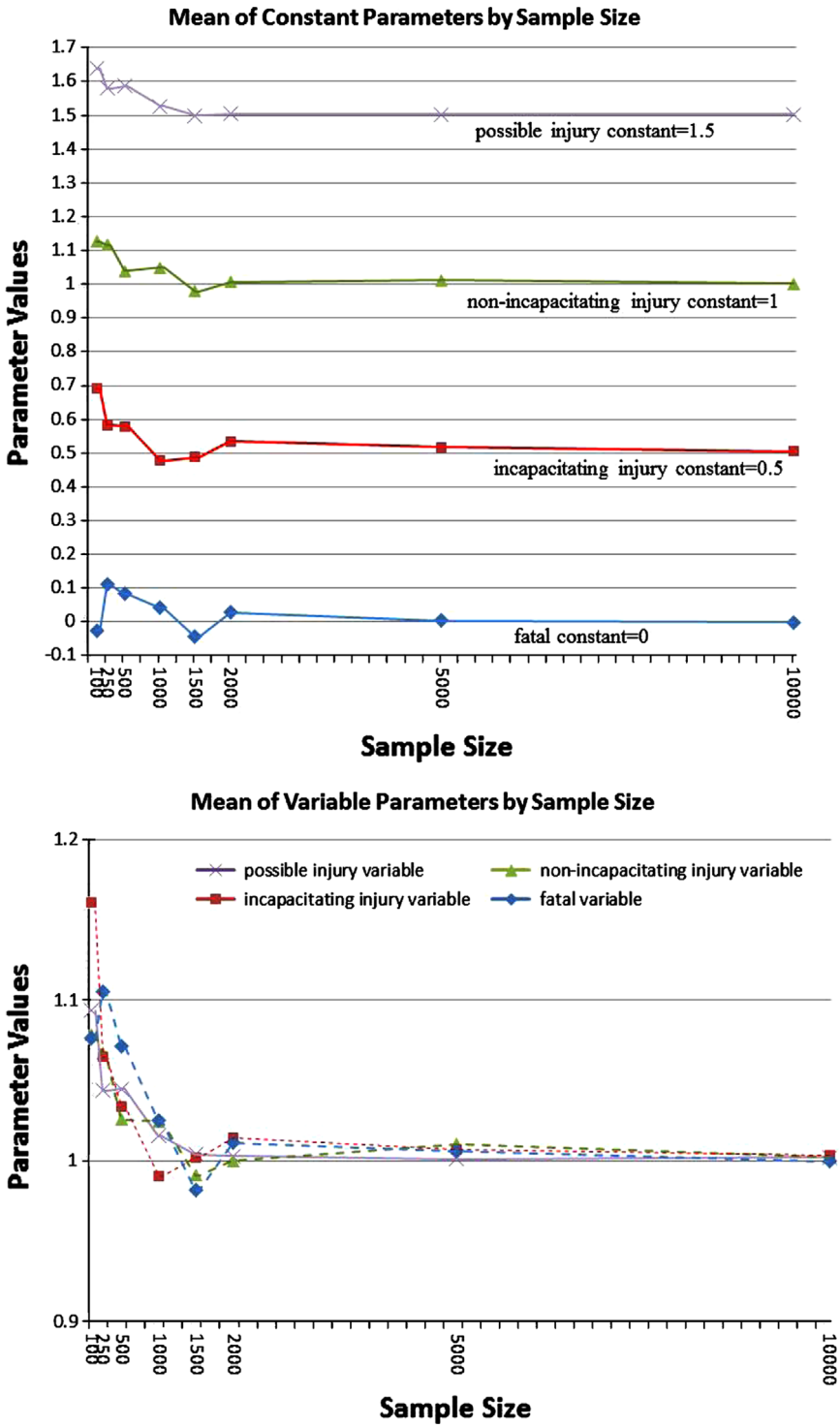

Fig. 2. Mean of the parameters by sample size for the multinomial logit model.

For the ordered probit model, as shown in Figs. 3 and 4, larger sample sizes lead to the narrower range for the 95% confidence interval for the parameters and closer value for the mean. As opposed to the multinomial logit model, the only difference is that for the ordered probit model, the stable point arrives at a smaller sample size, which is about half of that for the multinomial logit model (1000). In other words, as the sample size reaches 1000, the 95% confidence interval of parameters gets narrower and stable around the true value and the mean value is steadily close to the true value for each parameter. Similar to the multinomial logit model, the estimated mean value for all the variables appears to be biased for a sample size below 1000 observations.

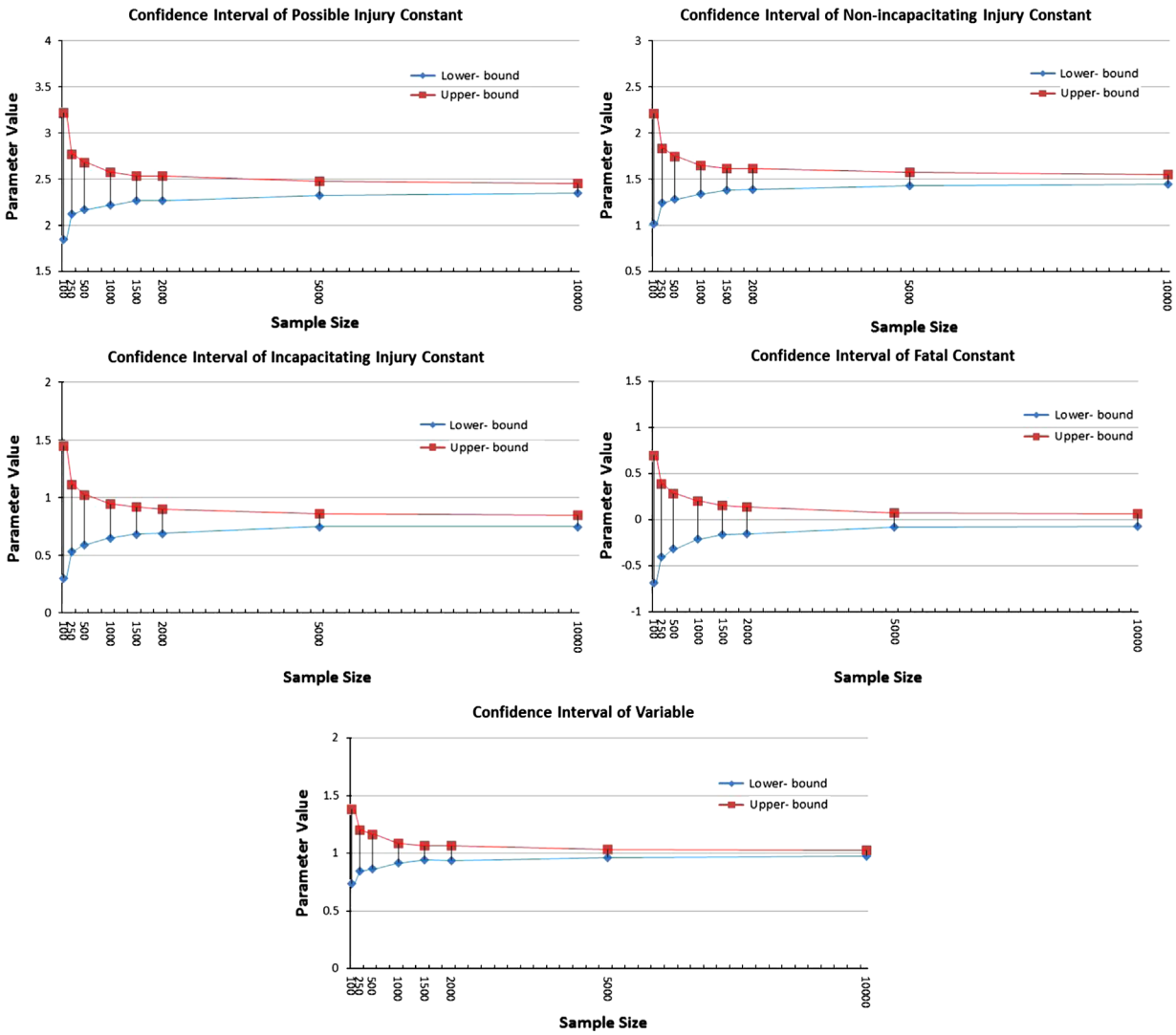

Fig. 3. Confidence intervals of the parameters by sample size for the ordered probit model.

For the mixed logit model, Figs. 5 and 6 show the relationships between both the 95% confidence intervals for the parameters and the mean value for each parameter as a function of the sample size. Very similar patterns as those observed above can be seen in these two figures. However, some differences can be noticed for the fatal variable parameter. Since the fatal variable parameter is random, it was found to be less stable both for the 95% confidence interval and the estimated mean value, especially for smaller sample sizes. In fact, the stable point for the fatal variable parameter is located around the 5000 observations mark (we use it as the stable point for the mixed logit model), which is the largest amongst the three models. Finally, it is anticipated that a larger sample size may be needed for the mixed logit model, if more random-parameters are introduced into the model.

To summarize, although the above results are based on simulated data, there are still a few findings that could be generalized in terms of sample size for the three models. Crash severity models with sample sizes below 1000 should not be estimated. In addition, the ordered probit model is the one that requires the least samples (>1000), the mixed logit model is the most demanding on samples (>5000), while the multinomial logit model requirements are located between the ordered probit and mixed logit models (>2000).

5.2. Analysis based on crash data

In the section above and for the sake of simplicity, we only included one variable which was assumed to be normally distributed. However, the crash severity data have a large amount of variation which might lead to different sample size requirements for the three models. Thus, we conducted further analyses using crash data described in Section 4. For this part, we

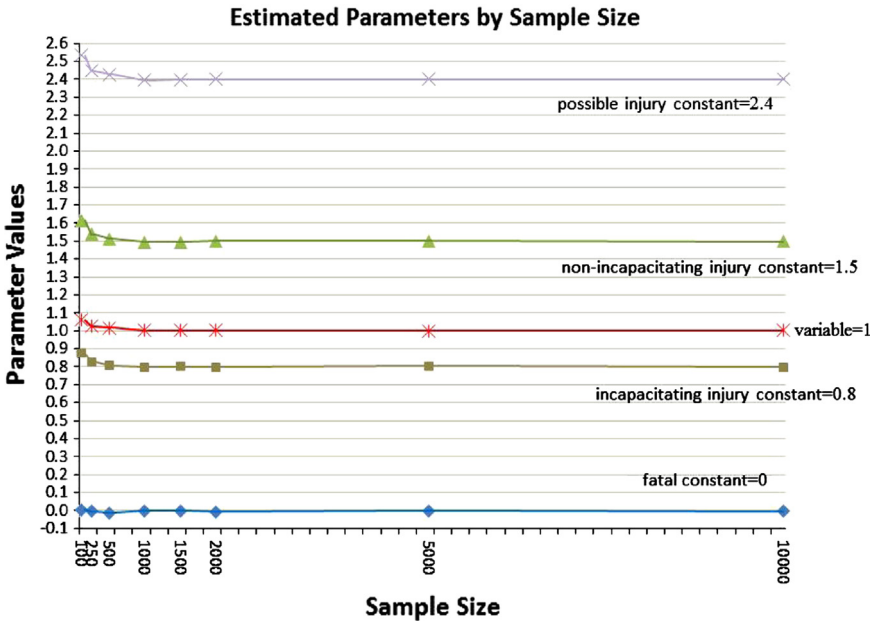

Fig. 4. Mean of the parameters by sample size for the ordered probit model.

set the models estimated from the full dataset as the baseline conditions (as estimated in Section 4). Then, the multinomial logit, mixed logit and ordered probit models were estimated using a stratified sampling method for different sampling sizes: 100, 500, 2000, 5000, 10,000, and 20,000 crashes. The stratified sampling method was used in order to keep the same proportion rates as those used for the full dataset: 45.3%, 20.1%, 22.2%, 9.4%, and 3.1% for severity no-injury, possible injury, non-incapacitating injury, incapacitating injury and fatal, respectively. In all, 30 random samples were selected for each sample size. We then compared the results with those calculated from the baseline conditions to get the value of bias, absolute-percentage-bias (APB) and root-mean-square-error (RMSE) for each parameter. Furthermore, the mean of APB, maximum of APB and total RMSE were estimated as a function of the sample size for each model.

Based on the 30 estimated models, for each parameter, the bias was calculated as $\text{Bias} = E(\hat{\beta}_r) - \beta_{\text{baseline}}$ (where r is the number of replications ($r=30$), β represents each parameter in the model, and $E(\hat{\beta}_r)$ is approximated by $\overline{\beta} = (1/r)\sum_{i=1}^{r}\hat{\beta}_i$). The APB was computed by dividing the absolute value of bias to the baseline value. The RMSE was calculated as $\text{RMSE} = \sqrt{\text{Bias}^2 + \text{Var}}$. Thus, the mean of the APB among all the parameters in a model could be calculated by taking the average of the APB values of all parameters. Furthermore, the maximum of APB was found by comparing the APB value of each parameter in a model. Finally, total RMSE could easily be attained by summing up the RMSE value of each parameter in a model. The results of the comparison analysis based on the three evaluation criteria described in the previous paragraph are summarized in Table 4.

Table 4
Three evaluation criteria by sample size for the three models*.

Sample size	Mean of absolute-percentage-bias (APB)			Max of absolute-percentage-bias (APB)			Total root-mean-square-error (RMSE)		
	MNL	ML	OP	MNL	ML	OP	MNL	ML	OP
100	5.50E+13	2.10E+11	143%	9.70E+14	2.90E+12	2.10E+01	7.40E+15	1.60E+13	20.7
500	2.00E+14	1.10E+04	25%	4.50E+15	1.10E+05	94%	1.30E+16	1.20E+06	4.5
2000	16%	26%	11%	45%	167%	40%	12.9	28.7	2.2
5000	9%	13%	5%	27%	52%	20%	7.6	13.7	1.2
10,000	4%	5%	4%	13%	13%	14%	4.7	8.7	0.7
20,000	2%	3%	2%	9%	21%	9%	1.9	3.4	0.4

* MNL: multinomial logit model; OP: ordered probit model; ML: mixed logit model.

From Table 4, we note the following results:

(1) As anticipated, all three models show the same tendency noted with the simulated data: the increase in sample size leads to the reduction in all three criteria (mean of APB, max of APB and total RMSE), improving the accuracy of model

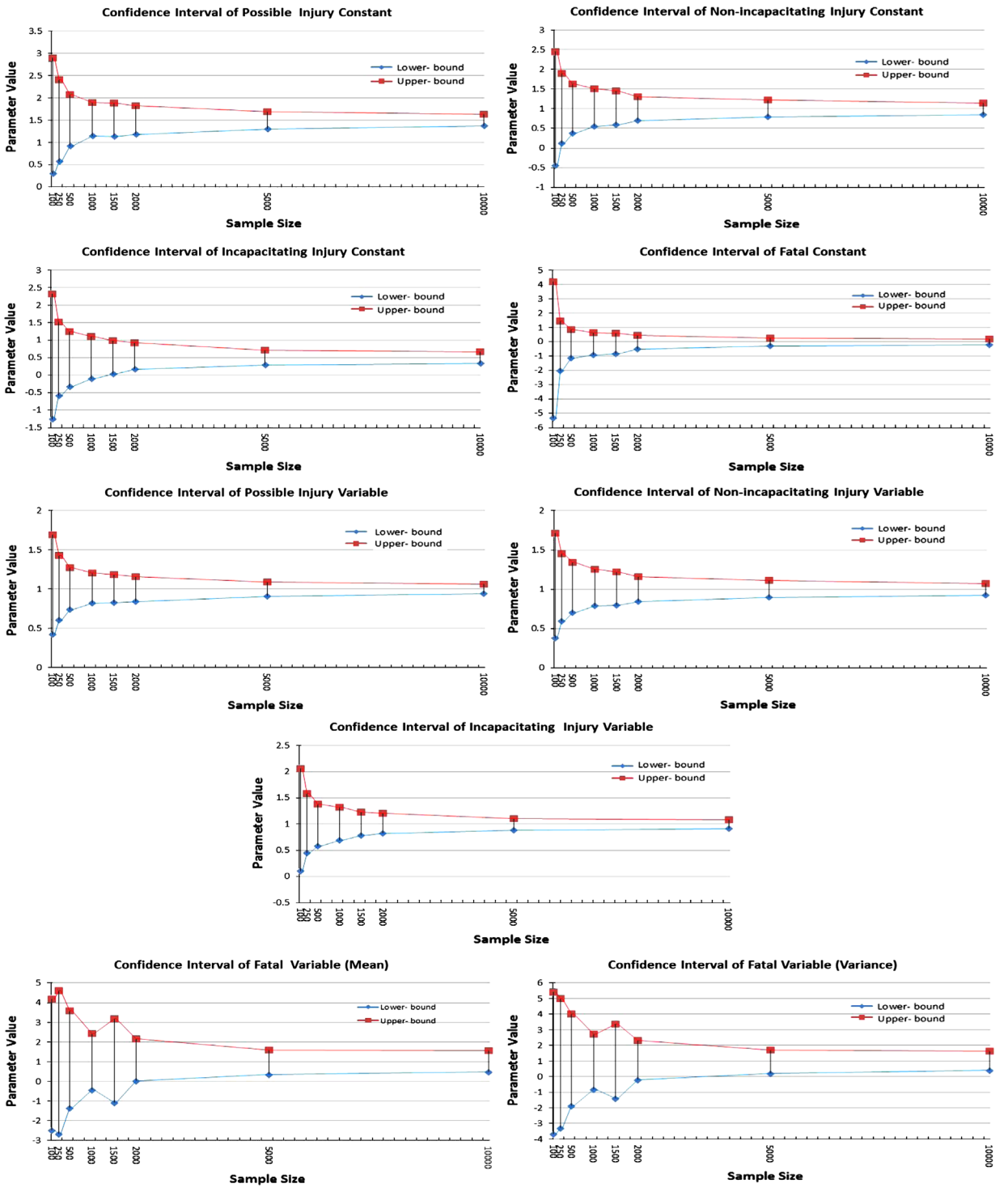

Fig. 5. Confidence intervals of the parameters by sample size for the mixed logit model.

estimation. Again, since maximum likelihood estimators are consistent, as the sample size approaches infinity (or becomes very large), the estimated parameters will have smaller standard errors and this leads to a better model performance (Washington et al., 2011).

(2) In terms of the values of all three criteria, the multinomial logit and mixed logit are more sensitive to small sample sizes than the ordered probit model and this is especially noticeable for the sample sizes equal to 100 and 500. Nonetheless, for a sample size below 500, all models perform poorly.

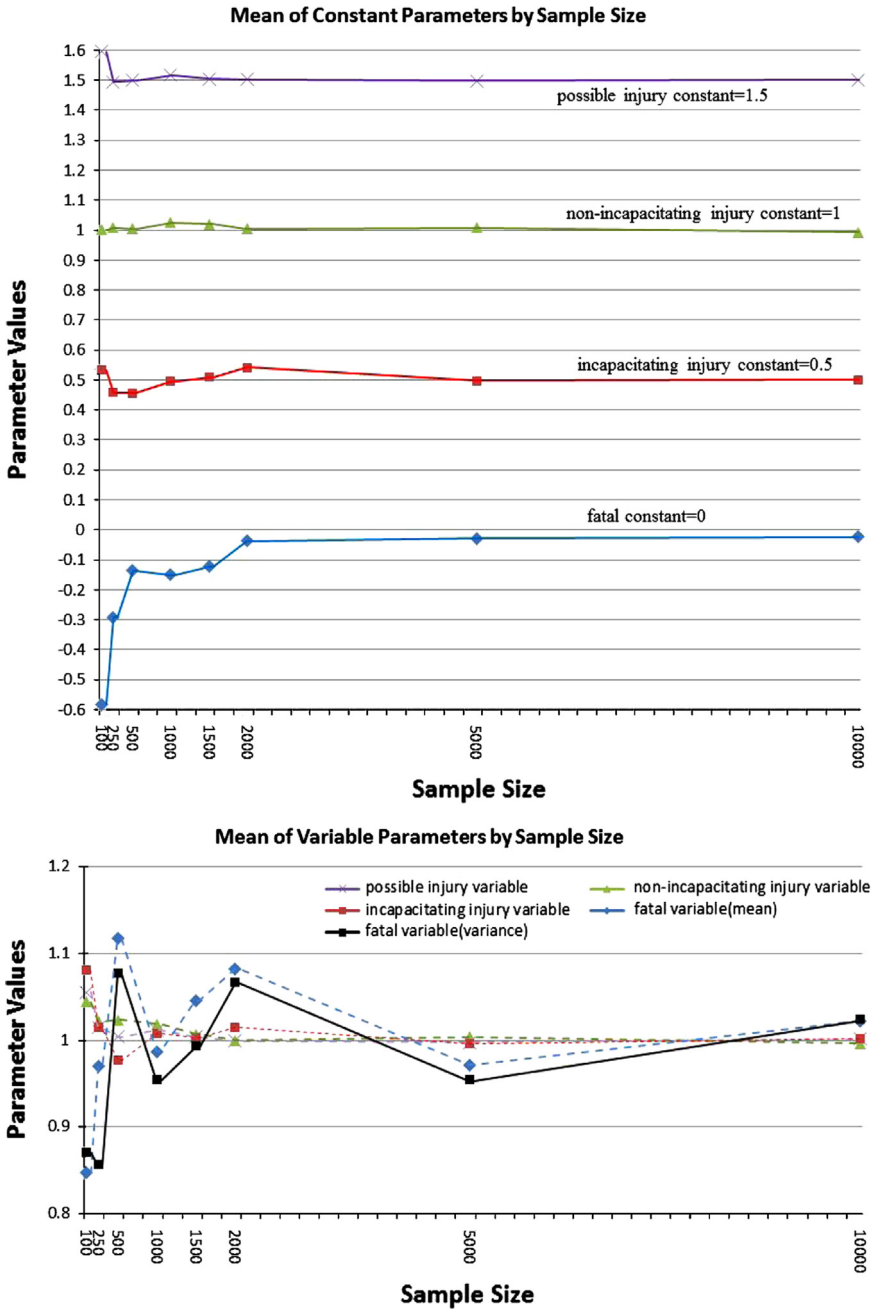

Fig. 6. Mean of the parameters by sample size for the mixed logit model.

(3) Similar to the results shown in the previous section, the mixed logit model needs a lot of data to lower the value of the three criteria. Even at 5000 observations, the mean of APB, max of APB and total RMSE for the mixed logit model is still twice as large as those for the multinomial logit model.

(4) According to the three criteria, the minimum sample size for the ordered probit, multinomial logit, and mixed logit models should be 2000, 5000 and 10,000, respectively. At that point, the estimated values become very close to the "true" values for all three criteria. In short, these findings are consistent with those found with the simulated data about which models are more affected by the small sample size problem. However, the minimum numbers are larger than the ones proposed in simulation. This may be partly explained by the large variability of crash data and the number of random samples running (30 for each sample size).

6. Conclusions and recommendations

There have been a lot of studies that have documented the application of crash severity models to explore the relationship between accident severity and its contributing factors such as driver characteristics, vehicle characteristics, roadway conditions, and road-environment 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 models. Similar to count data models, small data sets could significantly influence model performance. The objective of this study consisted in examining and quantifying the effects of different sample sizes on the performance of the three most commonly used crash severity models: the multinomial logit, ordered probit and mixed logit models. The objective of this study was accomplished by using a Monte-Carlo analysis based on simulated data and observed data. The sample size investigated varied between 100 and 10,000 observations.

Using 26,175 single-vehicle traffic accidents involving fixed-objects on rural two-way highways in Texas, it was first found that the mixed logit model has a better interpretive power than the multinomial logit model, while this latter model had superior interpretive power than the ordered probit model. On the other hand, the ordered probit model had a slightly better goodness-of-fit than that of the multinomial logit and mixed logit models, but the mixed logit model had a significant better fit than the multinomial logit model.

The results from the simulated data and random samples drawn from 26,175 crash records 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. Overall, the recommended absolute minimum numbers of observations for the ordered probit, multinomial logit, and mixed logit models are 1000, 2000 and 5000, respectively. Although those values are recommended guidelines, larger datasets should be sought, as demonstrated by the analysis using observed crash data (larger variability in the crash data or more randomness estimated in the mixed logit model). In order to minimize the bias produced by the insufficient sample size, the sequence of selecting a model among the three ones is the ordered probit model, multinomial logit model and mixed logit model as mentioned above. This study is a first step in the model comparison of the sample size on crash severity models. Further research is needed to generalize sample size requirements for developing the three models evaluated in this study, which may be partly dependent upon the characteristics of the data, as discussed in Savolainen et al. (2011). Finally, the same kind of research should be expanded to other crash severity models (e.g., the random parameters ordered probit model, finite mixture models, Markov switching models, etc.).

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34 (5), 597–603.

Abdel-Aty, M., Ekram, A.-A., Huang, H., Choi, K., 2011. A study on crashes related to visibility obstruction due to fog and smoke. Accident Analysis and Prevention 43 (5), 1730–1737.

Eluru, N., Bhat, C., Hensher, D., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis and Prevention 40 (3), 1033–1054.

Greene, W.H., 2007. LIMDEP User's Manual: Version 9.0. Econometric software, Plainview, NY.

Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. Journal of Safety Research 41 (4), 347–357.

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accident Analysis and Prevention 37 (5), 910–921.

Kim, J-K., Ulfarsson, G., Shankar, V., Kim, S., 2008. Age and pedestrian injury severity in motor-vehicle crashes: a heteroskedastic logit analysis. Accident Analysis and Prevention 40 (5), 1695–1702.

Lord, D., Bonneson, J., 2005. Calibration of predictive models for estimating the safety of ramp design configurations. Transportation Research Record 1908, 88–95.

Lord, D., 2006. Modeling motor vehicle crashes using poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis and Prevention 38 (4), 751–766.

Lord, D., Miranda-Moreno, L., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Safety Science 46 (5), 751–770.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.

Milton, J., 2006. Generalized Extreme Value and Mixed Logit Models: Empirical Applications to Vehicle Accident Severities. Civil and Environmental Engineering Department, University of Washington (Ph.D. dissertation, UMI no.: 3241933).

Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40 (1), 260–266.

Park, B.-J., Lord, D., Hart, J., 2010. Bias properties of bayesian statistics in finite mixture of negative regression models for crash data analysis. Accident Analysis and Prevention 42 (2), 741–749.

Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43 (5), 1666–1676.

Train, K.E., 2003. Discrete Choice Methods With Simulation. Cambridge University Press, New York, NY.

Washington, S., Karlaftis, M., Mannering, F., 2011. Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. Chapman and Hall/CRC, Boca Raton, FL.

Ye, F., 2011. Investigating the effects of sample size, model misspecification, and underreporting in crash data on three commonly used traffic crash severity models. Texas A&M University (Ph.D. dissertation, UMI no. 3471262).