Accident Analysis and Prevention 45 (2012) 646–653

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Analysis of crash severities using nested logit model—Accounting for the underreporting of crashes

Sunil Patil[a,*], Srinivas Reddy Geedipally[b], Dominique Lord[c]

[a] RAND Europe, Westbrook Center, Milton Road, Cambridge – CB4 1YG, UK
[b] Center for Transportation Safety, Texas Transportation Institute – The Texas A&M University System, College Station, TX – 77843-3135, USA
[c] Zachry Department of Civil Engineering, Texas A&M University, College Station, TX – 77843, USA

ARTICLE INFO

Article history:
Received 16 July 2011
Received in revised form 22 August 2011
Accepted 18 September 2011

Keywords:
Crash injury severity
Nested logit
Underreporting

ABSTRACT

Recent studies in the area of highway safety have demonstrated the usefulness of logit models for modeling crash injury severities. Use of these models enables one to identify and quantify the effects of factors that contribute to certain levels of severity. Most often, these models are estimated assuming equal probability of the occurrence for each injury severity level in the data. However, traffic crash data are generally characterized by underreporting, especially when crashes result in lower injury severity. Thus, the sample used for an analysis is often outcome-based, which can result in a biased estimation of model parameters. This is more of a problem when a nested logit model specification is used instead of a multinomial logit model and when true shares of the outcomes-injury severity levels in the population are not known (which is almost always the case). This study demonstrates an application of a recently proposed weighted conditional maximum likelihood estimator in tackling the problem of underreporting of crashes when using a nested logit model for crash severity analyses.

1. Introduction

Vehicle crash is one of the most common causes of death and injury around the world. This has resulted in the proclamation of the current decade (2011–2020) as the *Decade of Action for Road Safety* by the United Nations (United Nations, 2010). Highway safety professionals often aim at reducing the number and associated severity of traffic crashes through the analysis of existing reported data. Studies focusing on analyzing traffic crash severities aim at identifying and quantifying the effects of the factors, which affect different crash injury severities. Recently, researchers have used logit models for this kind of analyses.

Crash injury severities are often classified as follows: fatal, incapacitating, non-incapacitating, slight or possible injury and no injury/property damage only. The data used in these studies are often collected by the police from reported crashes. However, less severe crashes are often underreported due to various reasons which include avoidance of reporting by the driver(s) involved (Hauer and Hakkert, 1988; Elvik and Mysen, 1999; Blincoe et al., 2002). Thus, studies, which assume a random sampling strategy, are likely to result in producing biased results of parameter estimation (Savolainen et al., 2011). While sampling bias can be corrected when the population shares of these severity levels are known, it is very rare for crash data to obtain these true shares.

Recently, Bierlaire et al. (2008) have proposed a new estimator, which can account for outcome (choice)-based sampling when the population shares of the outcomes are unknown. This estimator is proposed for the model structures, such as the nested logit (NL) model, which belong to the wider generalized extreme value (GEV) family. The main objective of our study is to investigate if this estimator can be useful in addressing the crash underreporting problem when a NL model is used for crash injury severity analysis.

In the next section, we present a review of relevant literature. This is followed by the sections on methodology, details of data, results, and conclusions.

2. Literature review

Various model structures have been used to model crash injury severities. A detailed review and assessment of these models is presented in a recent paper (Savolainen et al., 2011). Many of the studies including those carried out by Shankar et al. (1996), Chang and Mannering (1998), Chang and Mannering (1999), Lee and Mannering (2002), Abdel-Aty and Abdelwahab (2004), Holdridge et al. (2005), Savolainen and Mannering (2007), Haleem and Abdel-Aty (2010), and Hu and Donnell (2010) have used the NL model specification to analyze the crash injury severity data.

* Corresponding author. Tel.: +44 1223 353 329; fax: +44 1223 358 845.
E-mail addresses: spatil@rand.org, patilnsunil@gmail.com (S. Patil), srinivas-g@ttimail.tamu.edu (S.R. Geedipally), d-lord@tamu.edu (D. Lord).

0001-4575/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2011.09.034

The NL and multinomial logit (MNL) model specifications are used in the crash injury severity data analysis for various reasons. These specifications are more flexible in terms of capturing the effects of independent variables when compared to the ordered response probit (OP) models (see Savolainen and Mannering, 2007). Also, as MNL and NL specifications belong to the GEV family, they are deterministic and do not need assumptions regarding the distributions of the parameters, as opposed to the mixed/random parameter logit (ML) model specification (McFadden, 1978).

Furthermore, the NL specification is less restrictive than the MNL model as it partly avoids reliance on the assumption of 'independence from irrelevant alternatives' (IIA) (see Train (2009) for a detailed discussion). While the MNL model offers great simplicity through its assumption of IIA, the simplicity often becomes a weakness. The IIA property of the MNL restricts the ratio of probabilities for any pair of alternatives to be independent of the existence and characteristics of other alternatives in a set of alternatives. This restriction implies that the introduction of a new alternative to the set will affect all other alternatives proportionately. IIA thus implies symmetry of alternatives and hence equal cross-elasticities, which are rarely present in observed data. In observed data some alternatives (severity levels) may be more closely related than others and for this reason IIA is not acceptable in variety of outcomes. The NL specification, however, offers more realistic substitution pattern among the severity levels, hence the NL is often preferred over the MNL specification. However, the NL models used in above mentioned studies are estimated without accounting for the possibility of underreporting or the sampling bias. This implicit assumption of random sampling may not always be true.

Many researchers have documented the presence of underreporting in crashes. These include: Hauer and Hakkert (1988), Hvoslef (1994), James and Kim (1996), Stutts and Hunter (1998), Aptel et al. (1999), Elvik and Mysen (1999), Alsop and Langley (2001), Cryer et al. (2001), Dhillon et al. (2001), Rosman (2001), Amoros et al. (2006), Hauer (2006), Tsui et al. (2009), and Savolainen et al. (2011). Underreporting of crashes can be attributed to several reasons.

In some cases, underreporting is a direct result of lack of reporting by individuals involved in crashes that result in minor or no injury (in an effort to avoid possible traffic citation and involvement of insurance company, which may result in an increase in the cost of vehicle insurance). Further, these data are most often collected by police reporting at the crash site. There is a possibility of inconsistency in the classification of a crash outcome into no injury or possible injury levels (see Savolainen and Mannering, 2007) and/or presence of threshold to record the crash only if the vehicle or property damages exceed a certain amount (Hauer, 2006). These factors will make the crash data an outcome-based/biased sample.

There is a lot of variation in the extent of underreporting identified by above studies and it depends on study location and severity levels. As per findings of Hauer and Hakkert (1988), approximately 20% of severe injuries, 50% of minor injuries, and up to 60% of no-injury crashes are not reported. A study by Elvik and Mysen (1999) reported underreporting rates of 30%, 75%, and 90% for serious, slight, and very slight injuries, respectively. Further, the National Highway Traffic Safety Administration (NHTSA, 2009) estimated that 25% of minor injury crashes and half of no-injury crashes may be unreported. On the other hand, Blincoe et al. (2002) have reported absence of underreporting for fatal crashes. Thus, there is a strong case to assume that crash data are often outcome-based samples and the exact shares of the outcomes (in the intended population of total crashes in the study area) are not known during model estimation.

In such instances when a conventional maximum likelihood estimator (MLE) is used, the MNL model will still produce unbiased estimates for all model parameters except the alternative specific constants (ASCs) (Cosslett, 1981a,b). However, the results of NL estimated using conventional MLE will be biased.

Yamamoto et al. (2008) studied the effects of underreporting when estimating OP and sequential binary probit models. They used a pseudo-likelihood function for the case with unknown population shares to examine the effects of underreporting on the parameter estimates. They concluded that ignoring the underreporting can result in significantly biased estimates of the effects of the explanatory variables and their elasticities. For example, they found that effects of environmental factors, safety restraint use and gender of driver can be over or under-estimated if conventional estimation techniques are used to estimate an OP model. However, as pointed in Ye and Lord (forthcoming), the validation and efficiency of the methods were not confirmed in this study.

Ye and Lord (forthcoming) studied the effects of underreporting on three commonly used crash severity model specifications, namely the MNL, ML and OP. Using simulated and observed (synthetic) data, Ye and Lord (forthcoming) concluded that estimated parameters of these models (which included the ASCs) are biased when underreporting is present. They also concluded that when some information regarding the extent of underreporting is available, using a weighted exogenous sample maximum likelihood estimator (WESMLE) produces better results than using the MLE.

In this study, we investigate the possibility of using a new estimator proposed by Bierlaire et al. (2008) for estimating NL models in the presence of underreporting and unknown population shares of different severity levels. As per our knowledge, this is the first instance of estimation of NL model for crash injury severities, which accounts for the underreporting of crashes. In next section, we describe the NL model specification and the estimator to account for underreporting of crashes.

3. Methodology

Discrete choice/outcome models such as the NL model have been successfully applied to analyze crash injury severity levels as a function of different covariates in many studies mentioned in Section 2. In these models, the crash injury severity level (fatal, incapacitating, etc.), which is a discrete outcome, is considered as the dependent variable. The probability of occurrence of these severity levels, given that a crash has occurred, is specified as a function of various roadway, driver, weather, vehicle, and crash characteristics. In this section, we present how a NL model for crash injury severities can be specified to account for underreporting, using the estimator of Bierlaire et al. (2008).

For a roadway segment n, an individual crash severity i among the given severities J is considered to be predicted if the crash severity likelihood function (U_{in}) is maximum for that particular severity. Each crash severity likelihood function, which is a dimensionless measure of the crash likelihood, is considered to have a deterministic component (V_i) and an error/random component (Eq. (1)).

$$U_i(\boldsymbol{x}, \boldsymbol{\beta}) = V_i(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon_i \tag{1}$$

where x is the vector containing independent variables, β is a vector of unknown parameters to be estimated which may include an ASC, and ε_i is a random term (here, we have ignored the segment-specific term n for the sake of simplicity). The random components ε_i are assumed to represent the unobserved and unaccountable factors involved in the determination of the outcome, whether these are derived from inherent randomness in the occurrence of a severity level at a given segment or caused by a lack of knowledge of this process on the part of the analyst. Specific assumptions on the distribution of ε_i yield to different specifications of models, such as logit, probit, etc. For a model from the GEV family the probability

of occurrence of an injury severity level i from the set of all severity levels $\mathcal{C}$ is given by Eq. (2).

$$P(i|\boldsymbol{x},\boldsymbol{\theta}) = \frac{\Lambda_i(\boldsymbol{x},\boldsymbol{\theta})}{\sum_{j\in\mathcal{C}}\Lambda_j(\boldsymbol{x},\boldsymbol{\theta})} \tag{2}$$

where

$$\Lambda_i(\boldsymbol{x},\boldsymbol{\theta}=(\boldsymbol{\beta};\boldsymbol{\gamma})) = e^{V_i(\boldsymbol{x},\boldsymbol{\beta})+\ln G_i(\left\langle e^{V_k(\boldsymbol{x},\boldsymbol{\beta})}\right\rangle_{k\in c};\boldsymbol{\gamma})+\ln S(i,\boldsymbol{\theta})} \tag{3}$$

and

$$G_i(\left\langle y_k\right\rangle_{k\in\mathcal{C}};\boldsymbol{\gamma}) = G_i(y_1,\ldots y_J;\boldsymbol{\gamma}) = \frac{dG}{dy_i}(y_1,\ldots y_J;\boldsymbol{\gamma}) \tag{4}$$

where $y_k = e^{V_k(x,\beta)}$. The function $G(\cdot)$ is called a μ-GEV-generating function, which depends on a set of unknown parameters in vector y. It is defined such that it satisfies the limit and signed derivative conditions given by McFadden (1978). $G(\cdot)$ is a homogeneous function of degree μ in $(y_1,\ldots y_J)$ and it gives rise to numerous model specifications from GEV family, including MNL and NL models. The term $S(i,\boldsymbol{\theta})$ represents the sampling probability for the roadway segment i and this term can contain unknown parameters to be estimated, given by θ (see, Bierlaire et al. (2008) for details).

For the NL model, the set $\mathcal{C}$ is partitioned into M mutually exclusive groups/nests and the function $G(\cdot)$ is specified as:

$$G(y;\boldsymbol{\gamma}=\{\mu,\mu_1,\mu_2\ldots\mu_M\}) = \sum_{m=1}^{M}\left(\sum_{j\in\mathcal{C}_m} y_j^{\mu_m}\right)^{\mu/\mu_m} \tag{5}$$

The ratio μ/μ_m, is commonly referred to as nest/inclusive value parameter and the scale parameter μ is usually normalized to one. An inclusive value parameter equal to one indicates that there is no correlation in the unobserved factors within the nest; hence, the model is not different than the standard MNL model, whereas a value of zero indicates perfect correlation among the different severity levels.

From Eqs. (3)–(5), it can be verified that for the MNL model ($\mu_m = 1$, $\mu/\mu_m = 1$) the term $\ln G_i(\cdot)$ vanishes and the term $\ln G_i(\cdot)$ is confounded with the ASC. Hence, in the presence of underreporting only the estimates of ASCs will be biased (shifted by $\ln S(\cdot)$) in a MNL model.

In the case of the NL model, however, the term $\ln G_i(\cdot)$ in Eq. (3) is not zero when there are more than one severity levels in a nest. Hence, both ASC and the sampling probability are identified. However, for the dummy nest with only one alternative, the term $\ln G_i(\cdot)$ is zero and hence ASC is shifted by $\ln S(\cdot)$ and only the sum is identified.

The parameters in the probability function can be estimated using the estimator proposed by Bierlaire et al. (2008), hereafter referred to as the 'new estimator' in this study. They can be obtained by solving

$$\max_{\beta,y,\omega}\sum_{n=1}^{N}\ln\frac{e^{V_{i_n}(\boldsymbol{x}_n,\boldsymbol{\beta})+\ln G_{i_n}(\langle e^{V_k(\boldsymbol{x}_n,\boldsymbol{\beta})}\rangle_{k\in\mathcal{C}};\boldsymbol{\gamma})+\omega_{i_n}}}{\sum_{j\in\mathcal{C}} e^{V_{jn}(\boldsymbol{x}_n,\boldsymbol{\beta})+\ln G_j(\langle e^{V_k(\boldsymbol{x}_n,\boldsymbol{\beta})}\rangle k\in\mathcal{C};\boldsymbol{\gamma})+\omega_j}} \tag{6}$$

where i_n is the observed severity level for the roadway segment n and $\omega_j = \ln S(j,\theta)$. The parameter ω_j, which accounts for underreporting can be estimated directly (when there are more than one severity levels in a nest). This estimator can also contain terms for weighting (sampling protocol) which we have omitted for simplicity and lack of relevance in this study.

We use this estimator to estimate a NL model for crash injury severities, which may have underreporting. The software package Biogeme (Bierlaire, 2003, 2007) was used for this purpose. Next, we describe the data used in this study along with the results of model estimation.

4. Data and results

We illustrate the application of the new estimator to account for the problem of underreporting of crashes using observed and synthetic datasets. A motorcycle crash dataset from Texas is used as the observed data and then synthetic datasets are derived from this dataset.

4.1. Observed data—motorcycle crash data in rural Texas

Data on crashes involving motorcycles that occurred on the Texas state highway system from 2003 through 2008 were obtained from the Texas Department of Transportation (TxDOT) crash records information system (CRIS). We used only crashes that occurred in rural areas. The dataset contains information on crash severity, crash type, roadway information, environmental condition, rider gender and age, vehicle age, and driving under alcohol influence. 'Crash severity' in this data refers to the most severely involved motorcyclist (motorcycle rider or passenger) in a particular crash. We used four crash severity levels: fatal, incapacitating, non-incapacitating, slight/no injury. The severity levels – slight injury and no injury were grouped as one category, because it is sometimes not possible to distinguish between these two categories due to variations in crash reporting (Savolainen and Mannering, 2007). The summary statistics for the variables used for model estimation are given in Table 1.

Geedipally et al. (forthcoming) have estimated MNL models of crash severity for this dataset after checking that NL model specifications were not justified for this data. However, they checked only those nest structures, which respected the ordinality of severity levels. However, ordinality of injury level does not hold true in all cases. Two important variables discussed in the literature that violate the ordinality assumption are seat belt usage and air-bag deployment. With these variables, the probability of both the fatal and no-injury severities decreases while increasing the probability of other severities. In this study, we tested (using conventional MLE estimator) all possible nest structures with the four severity levels, irrespective of the ordinality. We find that only three nest structures can be justified (μ_m should be significantly greater than one). We use one of these nest structures which yields the best fit to the dataset based on the log-likelihood value. In this nest structure, we grouped the severity levels into three nests as: Nest1-(fatal, no/slight injury), Nest2-(incapacitating), and Nest3-(non-incapacitating) as shown in Fig. 1. It should be noted that the nest structure forces choice of alternative for which of the ω parameter can be estimated, however, it is possible to account for underreporting in any/all of the alternatives as demonstrated in next section.

The severity level – fatal was considered as the base alternative (ASC = 0). Using this nest structure and the conventional MLE estimator, we estimated a NL model. Variables related to rider, roadway

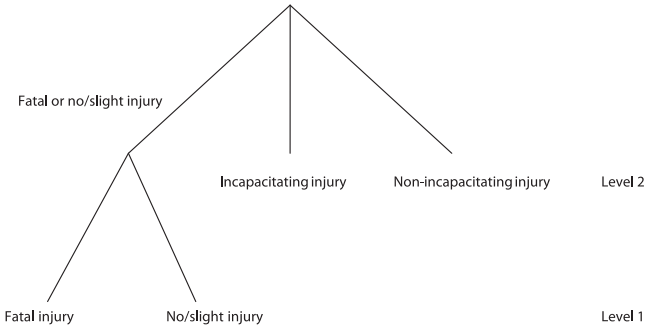

Fig. 1. Nest structure used for the analysis.

Table 1
Descriptive statistics for the Texas motorcycle crash data.

Variable	Value/details	Crash frequency	Percent
Intersection/intersection related (INT)	Yes (1)	1370	20.4%
	No (0)	5330	79.6%
Surface condition (SURF)	Good (1)	6353	94.8%
	Bad (0)	347	5.2%
Light condition (LGT)	Good (1)	5143	76.8%
	Bad (0)	1557	23.2%
Single vehicle crash (SV)	Yes (1)	4565	68.1%
	No (0)	2135	31.9%
Angular crash (ANG)	Yes (1)	481	7.2%
	No (0)	6219	92.8%
Same direction crash (SD)	Yes (1)	1214	18.1%
	No (0)	5486	81.9%
Divided road (DIV)	Yes (1)	1108	16.5%
	No (0)	5592	83.5%
Helmet used (HMT)	Yes (1)	3902	58.2%
	No (0)	2798	41.8%
Rider is intoxicated (ALC)	Yes (1)	269	4.0%
	No (0)	6431	96.0%
Motorcycle rider age (AGE)	Under 25 (U25)	1095	16.3%
	25–55 (O25)	4558	68.1%
	Over 55 (O55)	1047	15.6%
Fatal injury (FATAL)	–	576	8.6%
Incapacitating injury (INC)	–	2058	30.7%
Non-incapacitating injury (NINC)	–	2542	37.9%
No/slight injury (NOINJ)	–	1524	22.7%

segment geometry, weather, and crash characteristics were tested for inclusion in the model. Only those variables that were found to be significant at 10% level (t-test ≥ 1.64) were retained during the model estimation trials.

We then used the new estimator to estimate the same NL model. Note that, due to this nest structure, only one ω parameter in Eq. (6) is identified, in this case the one associated with severity level – no/slight injury (ω_NOINJ). This is due to the fact that second and third nests contain only one severity level ($\mu_2 = \mu_3 = 1$) and thus the ω parameters corresponding to the severity levels incapacitating and non-incapacitating are confounded with their ASCs. In the first nest, similar to the ASC related to severity level – fatal, the ω parameter is also constrained to zero.

Results of parameter estimation using both the conventional MLE estimator and the new estimator are given in Table 2. Notations b2, b3, b4 indicate that the coefficient belongs to the utility function of incapacitating, non-incapacitating or no-injury respectively. It can be observed from Table 2 that, the signs of all parameters (except the ASCs) are the same in both models. However, some variables lose significance in the model estimated using the new estimator. In general, the sign and magnitude of the coefficients for both the models in Table 2 are logical and consistent with previous research findings.

The parameter estimates in Table 2 indicate that given a crash has occurred, the probability of fatal crash increases when the rider is intoxicated or is older (above 55 years of age) or when the rider is not wearing a helmet. The probabilities of more severe injury levels increase (as compared to no-injury) when a rider involved is young (below 25 years of age).

In addition, the chance of fatality decreases for the motorcycles involved in an angular or same direction crash (when compared to other type of crashes such as head-on) or when it is a single-vehicle crash. The probability of high severity crashes decreases when the light conditions are good. Finally, divided highways decrease the chance of occurrence of high severity crashes. Although the coefficients related to variables – good surface condition (indicating higher probability of more severe crashes) and intersection-related crash (indicating lower probability of more severe crashes) appear counterintuitive, the effects could probably be explained by their strong correlation with vehicular speed.

Comparing the log-likelihood values for these two models, it can be observed that the model estimated using the new estimator provides a better fit to data (larger log-likelihood). A likelihood ratio test confirmed that the improvement in model fit is significant. The parameter, which accounts for underreporting (captures the sampling bias) ω_NOINJ was also found to be significantly different than zero. The standard errors of parameter estimation for most of the parameters except for the ASCs were found to be slightly lower with the new estimator.

Finally, while this estimation exercise demonstrates the use of new estimator; we do not have an idea about the real rate of underreporting in this data, which is often the case with police reported crash data. Hence, in order to compare these two estimators, a dataset with known rate of underreporting is needed. We used a simulated dataset derived from the observed crash data used in this section, by adding some noise so that the sampling process described in the next section could work (following Bierlaire et al.'s (2008), methodology).

4.2. Synthetic data

Using the above observed dataset, we generated a synthetic dataset as described here. We used observations related to each of the 6700 crashes to generate 100 synthetic observations (crashes), to obtain a total 'population' of 670,000 crashes. This will increase the size of synthetic data to be sufficient in drawing 100 subsamples from it. All the independent variables in this dataset were binary variables taking values 0 and 1, except for the age of a

Table 2
NL estimation results using conventional MLE and new estimator.

	Conventional MLE			New Estimator		
	Param.	Std. err.	*t*-test	Param.	Std. Err.	actual parameter estimated
Fatal injury (FATAL)						
ASC_FATAL (fixed)	0			0		
ω_FATAL (fixed)	n/a			0		
Incapacitating Injury (INC)						
ASC_INC	0.164	0.131	1.25	−1.760	0.659	−2.66 (ASC_INC + ω_INC − ω_FATAL)
b2ALC	−0.683	0.194	−3.52	−0.526	0.187	−2.82
b2ANG	0.422	0.190	2.22	0.225	0.189	1.19
b2AGE_O55	−0.192	0.110	−1.74	−0.147	0.102	−1.45
b2SD	0.394	0.178	2.21	0.187	0.179	1.04
b2SV	0.893	0.180	4.96	0.669	0.182	3.67
Non-incapacitating injury (NINC)						
ASC_NINC	−0.885	0.158	−5.62	−2.800	0.652	−4.29 (ASC_NINC + ω_NINC − ω_FATAL)
b3ALC	−0.929	0.199	−4.68	−0.786	0.191	−4.12
b3ANG	1.280	0.202	6.33	1.090	0.201	5.42
b3DIV	0.283	0.078	3.62	0.276	0.077	3.59
b3AGE_O55	−0.353	0.109	−3.23	−0.311	0.101	−3.07
b3HMT	0.312	0.058	5.35	0.298	0.058	5.11
b3LGT	0.263	0.066	4.00	0.257	0.065	3.95
b3SD	1.210	0.192	6.31	1.010	0.193	5.25
b3SV	1.840	0.194	9.47	1.620	0.195	8.31
No/slight injury (NOINJ)						
ASC_NOINJ	−0.380	0.159	−2.39	3.280	0.668	4.92
ω_NOINJ (shifted by ω_FATAL)	n/a			−6.030	0.999	−6.04 (ω_NOINJ − ω_FATAL)
b4ALC	−0.609	0.194	−3.14	−0.638	0.166	−3.85
b4ANG	0.642	0.191	3.36	0.683	0.165	4.13
b4DIV	0.287	0.086	3.33	0.324	0.083	3.89
b4AGE_O55	−0.321	0.102	−3.36	−0.343	0.092	−3.73
b4AGE_U25	0.156	0.068	2.31	0.191	0.071	2.70
b4HMT	0.364	0.069	5.25	0.414	0.064	6.46
b4INT	0.198	0.065	3.04	0.237	0.069	3.44
b4LGT	0.175	0.058	3.03	0.205	0.065	3.15
b4SD	0.851	0.215	3.96	0.931	0.162	5.74
b4SURF	−0.266	0.104	−2.55	−0.303	0.113	−2.69
b4SV	1.020	0.241	4.22	1.110	0.173	6.42
NEST1 (μ_1)	1.75	0.369	2.03	1.62	0.206	3.01
Log-likelihood		−8332.2				−8329.4
Number of observations		6700				6700

rider. For each binary variable, in order to add noise, we randomly changed its value from 0 to 1 or 1 to 0 in 40% of these 670,000 values. For the rider age variable, the value was generated by draw from a normal distribution N (μ, σ^2), where μ was the value in the original database and $\sigma = 0.05\,\mu$. This added noise is helpful in making sure that the subsamples drawn from synthetic population are not very similar to each other and to original observed data; however, these samples should still possess qualities similar to a real crash data.

A crash severity level was simulated for each of these crashes in the population using the NL model in Table 2 (estimated using conventional MLE). Thus, the 'true' values of the parameters for this exercise are considered to be known which were estimated from the observed data. Using this population of 670,000, we extracted 100 samples, each sample containing 6757 crashes assuming the underreporting rates as given in Table 3. Where, R_g is the probability for crash in a severity level g to be included in the sample. These under-reporting rates were chosen to represent one of the possible combinations identified from the rates reported in the studies mentioned in Section 2.

A model was then estimated for each of these 100 samples using the conventional MLE and the new estimator. Tables 4 and 5 summarize the results about the empirical distribution of parameter estimates and biases. Where,

- true – the 'true' value θ_k^* of parameter estimate (from Table 2),
- mean – the mean $\widehat{\theta}_k$,
- stdev – the std. dev. of 100 estimated parameters $\hat{\sigma}_k$,
- ttest – the ratio (mean – true)/std dev – to test if the mean is significantly different from the true value,
- min – the lowest value,
- max – the highest value,
- p5 – the 5th percentile,
- p95 – the 95th percentile,
- lowBound – $\theta_k^* - \Phi^{-1}(1\text{–}0.125)^* \hat{\sigma}_k$,
- upBound – $\theta_k^* + \Phi^{-1}(1\text{–}0.125)^* \hat{\sigma}_k$,
- count: the two-side 0.125 empirical coverage probability – % of estimates lying in the interval [lowBound, upBound]
- bias – $\widehat{\theta}_k - \theta_k^*$
- RMSE – $\sqrt{\text{bias}^2 + var^2}$, where var is variance of estimated parameter values

It can be observed from Table 4 (the conventional MLE estimator) that the parameter estimation results, except the ASCs, are not significantly different at the 5% level (t-value > 1.96) from the true values, however, the empirical coverage probabilities are much below 75% for most of the parameters (presented in bold typeface).

Table 3
Details of sampling scheme.

Severity	Synthetic population	Sample	% In syn. population	% In samples	Assumed rate of underreporting	R_g	$\omega = \ln(R_g)$ (shifted by ω_FATAL)
FATAL	56,619	944	8.5%	14.0%	0%	0.01667	−4.09 (0)
INC	149,924	2499	22.4%	37.0%	0%	0.01667	−4.09 (0)
NINC	298,747	2490	44.6%	36.9%	50%	0.00833	−4.79 (−0.69)
NOINJ	164,710	824	24.6%	12.2%	70%	0.00500	−5.30 (−1.2)
Total	670,000	6757	100%	100%	n/a	n/a	n/a

Table 4
Nested logit model on synthetic data estimated using conventional MLE.

	True	Count	lowBound	upBound	min	max	p5	p95	mean	stdev	ttest	bias	RMSE
ASC_INC	0.1643	**9**	0.0792	0.2495	0.1590	0.5490	0.2250	0.4690	0.3473	0.0740	−2.47	0.18	0.18
ASC_NINC	−0.8852	**0**	−1.0046	−0.7659	−1.6900	−1.1500	−1.5700	−1.2200	−1.4023	0.1038	4.98	−0.52	0.52
ASC_NOINJ	−0.3803	**1**	−0.5884	−0.1722	−1.4800	−0.5680	−1.2900	−0.6780	−0.9885	0.1809	3.36	−0.61	0.61
b2ALC	−0.6826	**66**	−0.7862	−0.5790	−0.9830	−0.5290	−0.9170	−0.6060	−0.7516	0.0901	0.77	−0.07	0.07
b3ALC	−0.9291	**56**	−1.0282	−0.8299	−1.2100	−0.7390	−1.1500	−0.8670	−1.0017	0.0862	0.84	−0.07	0.07
b4ALC	−0.6095	71	−0.7426	−0.4764	−0.8690	−0.3020	−0.8040	−0.3950	−0.5799	0.1157	−0.26	0.03	0.03
b2ANG	0.4218	**62**	0.3181	0.5255	0.2480	0.7340	0.3450	0.6610	0.4929	0.0902	−0.79	0.07	0.07
b3ANG	1.2791	**59**	1.1722	1.3860	1.1000	1.5600	1.2100	1.5100	1.3534	0.0929	−0.80	0.07	0.07
b4ANG	0.6418	70	0.5058	0.7777	0.3020	0.8660	0.4330	0.8110	0.6071	0.1182	0.29	−0.03	0.04
b3DIV	0.2830	74	0.2226	0.3433	0.1690	0.4000	0.1990	0.3790	0.2905	0.0525	−0.14	0.01	0.01
b4DIV	0.2871	**69**	0.2146	0.3595	0.1300	0.4220	0.1440	0.3720	0.2592	0.0630	0.44	−0.03	0.03
b2AGE_O55	−0.1915	**67**	−0.2935	−0.0896	−0.4530	0.0020	−0.3750	−0.0991	−0.2394	0.0886	0.54	−0.05	0.05
b3AGE_O55	−0.3534	**66**	−0.4706	−0.2363	−0.6070	−0.1190	−0.5900	−0.2290	−0.4026	0.1018	0.48	−0.05	0.05
b4AGE_O55	−0.3212	77	−0.4295	−0.2128	−0.6260	−0.0643	−0.4630	−0.1430	−0.2976	0.0942	−0.25	0.02	0.03
b4AGE_U25	0.1564	72	0.0749	0.2378	−0.0438	0.2980	0.0176	0.2740	0.1281	0.0708	0.40	−0.03	0.03
b3HMT	0.3120	71	0.2433	0.3808	0.1810	0.4810	0.2300	0.4410	0.3291	0.0598	−0.29	0.02	0.02
b4HMT	0.3639	**67**	0.2880	0.4399	0.1490	0.5200	0.2210	0.4200	0.3210	0.0660	0.65	−0.04	0.04
b4INT	0.1984	76	0.1397	0.2570	0.0405	0.3000	0.0709	0.2560	0.1822	0.0510	0.32	−0.02	0.02
b3LGT	0.2628	78	0.1829	0.3427	0.0986	0.4750	0.1570	0.3740	0.2656	0.0694	−0.04	0.00	0.01
b4LGT	0.1752	**66**	0.1104	0.2399	0.0605	0.3140	0.0782	0.2560	0.1572	0.0563	0.32	−0.02	0.02
b2SD	0.3935	**49**	0.2694	0.5177	0.2130	0.7710	0.3240	0.6870	0.5007	0.1079	−0.99	0.11	0.11
b3SD	1.2125	**59**	1.0734	1.3516	1.0400	1.6800	1.1200	1.5100	1.3200	0.1210	−0.89	0.11	0.11
b4SD	0.8512	72	0.6615	1.0409	0.4550	1.2400	0.5250	1.0800	0.8198	0.1649	0.19	−0.03	0.04
b4SURF	−0.2663	76	−0.3429	−0.1897	−0.4070	−0.0814	−0.3440	−0.1270	−0.2514	0.0666	−0.22	0.01	0.02
b2SV	0.8930	**52**	0.7658	1.0202	0.7900	1.3200	0.8500	1.2100	1.0275	0.1106	−1.22	0.13	0.14
b3SV	1.8399	**54**	1.6997	1.9802	1.7000	2.2300	1.7800	2.1700	1.9695	0.1219	−1.06	0.13	0.13
b4SV	1.0157	71	0.8174	1.2140	0.5110	1.3000	0.6340	1.2500	0.9578	0.1724	0.34	−0.06	0.07
NEST1	1.7485	79	1.2880	2.2091	1.3700	3.6100	1.4500	2.7100	1.9657	0.4004	−0.54	0.22	0.27

Table 5
Nested logit model on synthetic data estimated using the new estimator.

	True	Count	lowBound	upBound	min	max	p5	p95	mean	stdev	ttest	Bias	RMSE
ASC_INC	0.1643	80	−0.0471	0.3757	−0.4720	0.5060	−0.2500	0.3840	0.1206	0.1838	0.24	−0.04	0.04
ASC_NINC +ω_NINC(shifted)	−1.5786	79	−1.8098	−1.3474	−2.3200	−1.2400	−2.0100	−1.3300	−1.6238	0.2010	0.23	−0.05	0.05
ASC_ NOINJ	−0.3803	83	−0.8714	0.1108	−1.1000	1.4900	−1.0300	0.4710	−0.3205	0.4270	−0.14	0.06	0.19
ω_NOINJ(shifted)	−1.2038	82	−2.1212	−0.2864	−3.8600	2.1800	−2.7200	−0.1200	−1.2416	0.7975	0.05	−0.04	0.64
b2ALC	−0.6826	73	−0.7958	−0.5694	−0.9650	−0.4720	−0.8800	−0.5200	−0.6644	0.0984	−0.19	0.02	0.02
b3ALC	−0.9291	76	−1.0316	−0.8265	−1.2400	−0.7200	−1.0800	−0.7800	−0.9200	0.0892	−0.10	0.01	0.01
b4ALC	−0.6095	75	−0.7346	−0.4844	−0.8340	−0.2200	−0.7820	−0.4110	−0.5946	0.1088	−0.14	0.01	0.02
b2ANG	0.4218	76	0.3037	0.5399	0.1010	0.6540	0.2440	0.5930	0.4048	0.1027	0.17	−0.02	0.02
b3ANG	1.2791	77	1.1645	1.3937	0.9590	1.6000	1.1200	1.4400	1.2697	0.0997	0.09	−0.01	0.01
b4ANG	0.6418	75	0.5148	0.7688	0.2170	0.8600	0.4660	0.7870	0.6195	0.1104	0.20	−0.02	0.03
b3DIV	0.2830	73	0.2216	0.3443	0.1660	0.3990	0.1950	0.3720	0.2810	0.0533	0.04	0.00	0.00
b4DIV	0.2871	**69**	0.2051	0.3691	0.0702	0.4340	0.1590	0.4110	0.2823	0.0713	0.07	0.00	0.01
b2AGE_O55	−0.1915	71	−0.2958	−0.0873	−0.3870	0.0866	−0.3490	−0.0486	−0.1929	0.0906	0.01	0.00	0.01
b3AGE_O55	−0.3534	76	−0.4692	−0.2377	−0.5890	−0.1130	−0.5440	−0.2010	−0.3582	0.1006	0.05	0.00	0.01
b4AGE_O55	−0.3212	78	−0.4303	−0.2121	−0.6150	−0.0744	−0.4840	−0.1470	−0.3046	0.0949	−0.18	0.02	0.02
b4AGE_U25	0.1564	71	0.0705	0.2422	−0.0556	0.3140	0.0162	0.2830	0.1381	0.0747	0.24	−0.02	0.02
b3HMT	0.3120	75	0.2418	0.3822	0.1540	0.4730	0.2240	0.4260	0.3168	0.0610	−0.08	0.00	0.01
b4HMT	0.3639	73	0.2810	0.4469	0.1110	0.5400	0.2360	0.4640	0.3470	0.0721	0.23	−0.02	0.02
b4INT	0.1984	80	0.1323	0.2645	0.0444	0.3170	0.0775	0.2890	0.1981	0.0574	0.00	0.00	0.00
b3LGT	0.2628	79	0.1837	0.3419	0.0961	0.4700	0.1530	0.3680	0.2602	0.0687	0.04	0.00	0.01
b4LGT	0.1752	74	0.1042	0.2461	0.0331	0.3250	0.0844	0.2920	0.1728	0.0617	0.04	0.00	0.00
b2SD	0.3935	80	0.2587	0.5283	0.0240	0.7720	0.1760	0.5700	0.3807	0.1172	0.11	−0.01	0.02
b3SD	1.2125	80	1.0694	1.3556	0.9060	1.6000	0.9990	1.4300	1.2071	0.1244	0.04	−0.01	0.02
b4SD	0.8512	75	0.6785	1.0238	0.3050	1.1600	0.5670	1.0700	0.8385	0.1501	0.08	−0.01	0.03
b4SURF	−0.2663	73	−0.3523	−0.1803	−0.4830	−0.0832	−0.3930	−0.1320	−0.2733	0.0748	0.09	−0.01	0.01
b2SV	0.8930	74	0.7318	1.0541	0.5220	1.2600	0.6670	1.1100	0.8821	0.1401	0.08	−0.01	0.02
b3SV	1.8399	82	1.6693	2.0105	1.4900	2.1700	1.5500	2.1000	1.8331	0.1483	0.05	−0.01	0.02
b4SV	1.0157	83	0.8299	1.2015	0.3250	1.3200	0.6760	1.2600	0.9844	0.1615	0.19	−0.03	0.04
NEST1	1.7485	91	1.2021	2.2950	1.3400	5.4900	1.4700	2.5100	1.8962	0.4750	−0.31	0.15	0.27

The quality of estimates in Table 5 (the new estimator) seems to be much better, where none of the parameters estimated is significantly different from the true value. The empirical coverage probabilities for almost all the estimates are close to 75%. Even the values of bias and RMSE are lower than those in Table 4. Note that, forcing ω_FATAL = 0 will cause estimate of ω_NOINJ to be shifted, hence estimate of ω_NOINJ should be compared to ω_NOINJ − ω_FATAL ($-5.30 + 4.09 = -1.2$). Thus, ω_NOINJ is also correctly estimated. Also, since the severity levels INC and NINC are alone in their nests, the estimated values of their ASCs will be confounded with respective ω parameters. Hence, the corresponding estimates should be compared to ASC_INC(true value) + ω_INC − ω_FATAL ($0.16 - 4.09 + 4.09 = 0.16$) and with ASC_NINC(true value) + ω_NINC − ω_FATAL ($-0.885 - 4.79 + 4.09 = -1.58$). It can be observed that these ASCs are also correctly estimated.

We repeated this simulation experiment using the under-reporting rates of 5%, 10%, 50% and 70% for FATAL, INC, NINC and NOINJ severities. For this new design too, the new estimator provided better results compared to that of MLE (we have not presented the results of this design in this paper to avoid repetition). However, it was observed that the MLE estimates with this new design were slightly better than the MLE estimates in Table 4. Thus, the relative rates of under-reporting in all severities play an important role in the quality of estimates. It was also demonstrated that the new estimator can account for underreporting in all of the severities together.

Thus, the new estimator seems to be useful in addressing the underreporting issue found in the crash data. A minor drawback about using this estimator is that, it is not currently available in most of the estimation packages, however, the estimation package used in this study is freely available. Also, while the presence of underreporting can be detected through significant estimation of the ω parameter(s), it's not possible to get an estimate of extent of the under-reporting.

It should also be noted that analysis and methodology in this study will still retain the inherent drawbacks of using the NL model specification over more flexible model specifications such as the mixed logit model. For example, NL model specification cannot be applied to panel data (to account for temporal correlations in the unobserved factors for same segment) and it cannot relax assumptions related to homogeneity of the effect corresponding to a factor due to unobserved segment related characteristics.

5. Conclusions

The underreporting associated with crashes is one of the most important challenges in crash data analysis. This is even more critical when crash severity levels are modeled using a NL model specification. Unlike the MNL model specification where all the parameter estimates except the ASCs are unbiased in the presence of underreporting, the estimated parameters with the NL model will be biased. Previous studies have proposed using estimators that use some information about the extent of underreporting or sampling bias. However, it is rarely the case that an accurate extent of underreporting is known with certainty, especially with the low injury severity crashes. So far, the studies using the NL model for crash injury severity analysis have not accommodated underreporting.

Using a new estimator proposed by Bierlaire et al. (2008) on observed and synthetic data, we have demonstrated that this estimator can be useful in studies using the NL model for analyzing crash data with underreporting. We observe that the quality of estimates is improved with the new estimator. The parameter estimates were found to have intuitive signs and magnitudes. Hence, this estimator can be preferred for estimation of NL models for crash severities. However, more studies are needed to confirm the efficiency of the estimator for large-scale crash data analysis.

This research can also be extended to study the change in parameter elasticity values, wider range of underreporting values and the effect of different nest structures. Further, this estimator is indeed proposed for all the GEV family models, hence any future studies, which may also involve advanced crash injury severity models from the GEV family (e.g. cross-nested logit model) can be benefited by use of this estimator.

Acknowledgement

We would like to thank two anonymous reviewers for their useful suggestions and comments to improve the paper.

References

Abdel-Aty, M., Abdelwahab, H., 2004. Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure. Accident Analysis and Prevention 36 (3), 447–456.
Alsop, J., Langley, J., 2001. Under-reporting of motor-vehicle traffic crash victims in New Zealand. Accident Analysis and Prevention 33 (3), 353–359.
Amoros, E., Martin, J.-L., Laumon, B., 2006. Under-reporting of road crash casualties in France. Accident Analysis and Prevention 38, 627–635.
Aptel, I., Salmi, L.R., Masson, F., Bourdé, A., Henrion, G., Erny, P., 1999. Road accident statistics: discrepancies between police and hospital data in a French island. Accident Analysis and Prevention 31, 101–108.
Bierlaire, M., 2003. BIOGEME: a free package for the estimation of discrete choice models. In: Proceedings of the Third Swiss Transportation Research Conference, Ascona, Switzerland., www.strc.ch.
Bierlaire, M., 2007. An introduction to BIOGEME 1.5 (http://biogeme.epfl.ch/).
Bierlaire, M., Bolduc, D., McFadden, D., 2008. The estimation of generalized extreme value models from choice-based samples. Transportation Research Part B 42, 381–394.
Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., Spicer, R., 2002. The economic impact of motor vehicle crashes. NHTSA Technical Report, Washington, DC.
Chang, L.-Y., Mannering, F., 1998. Predicting vehicle occupancies from accident data: an accident severity approach. Transportation Research Record 1635, 93–104.
Chang, L.-Y., Mannering, F., 1999. Analysis of injury severity and vehicle occupancy in truck and non-truck-involved accidents. Accident Analysis and Prevention 31 (5), 579–592.
Cosslett, S.R., 1981a. Efficient estimation of discrete-choice methods. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Choice Data with Econometric Applications. MIT Press, Cambridge, MA, pp. 51–111.
Cosslett, S.R., 1981b. MLE for choice-based samples. Econometrica 49, 1289–1316.
Cryer, P.C., Westrup, S., Cook, A.C., Ashwell, V., Bridger, P., Clarke, C., 2001. Investigation of bias after data linkage of hospital admission data to police road traffic crash reports. Injury Prevention 7 (3), 234–241.
Dhillon, P.K., Lightstone, A.S., Peek-Asa, C., Kraus, J.F., 2001. Assessment of hospital and police ascertainment of automobile versus childhood pedestrian and bicyclist collisions. Accident Analysis and Prevention 33 (4), 529–537.
Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting; meta-analysis of studies made in 13 Countries. Transportation Research Record 1665, 133–140.
Geedipally, S., Turner, P., Patil, S. An analysis of motorcycle crashes in Texas using a multinomial logit model. Transportation Research Record (forthcoming). Paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C.
Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. Journal of Safety Research, doi:10.1016/j.jsr.2010.04.006.
Hauer, E., Hakkert, A., 1988. Extent and some implications of incomplete accident reporting. Transportation Research Record 1185, 1–10.
Hauer, E., 2006. The frequency-severity indeterminacy. Accident Analysis and Prevention 38, 78–83.
Holdridge, J., Shankar, V., Ulfarsson, G., 2005. The crash severity impacts of fixed roadside objects. Journal of Safety Research 36 (2), 139–147.
Hu, W., Donnell, E.T., 2010. Median barrier crash severity: some new insights. Accident Analysis and Prevention 42 (6), 1697–1704.
Hvoslef, H., 1994. Under-Reporting of Road Traffic Accidents Recorded by the Police at the International Level. Public Roads Administration, Norway.
James, J.L., Kim, K.E., 1996. Restraint use by children involved in crashes in Hawaii, 1986–1991. Transportation Research Record 1560, 8–11.
Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of runoff-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34 (2), 149–161.
McFadden, D., 1978. Modelling the choice of residential location. In: Karlquist, A., et al. (Eds.), Spatial Interaction Theory and Residential Location. North-Holland, Amsterdam, pp. 75–96.

National Highway Traffic Safety Administration, 2009. Traffic Safety Facts: Motorcycles, DOT HS 811 159. National Highway Traffic Safety Administration, Washington, DC.
Rosman, D.L., 2001. The Western Australian road injury database (1987–1996): ten years of linked police, hospital and death records of road crashes and injuries. Accident Analysis and Prevention 33 (1), 81–88.
Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39 (5), 955–963.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The Statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43 (5), 1666–1676.
Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accident Analysis and Prevention 28 (3), 391–401.
Stutts, J., Hunter, W., 1998. Police reporting of pedestrians and bicyclists treated in hospital emergency rooms. Transportation Research Record 1635, 88–92.
Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press, United Kingdom, pp. 45–47.
Tsui, K.L., So, F.L., Sze, N.N., Wong, S.C., Leung, T.F., 2009. Misclassification of injury severity among road casualties in police reports. Accident Analysis and Prevention 41 (1), 84–89.
United Nations, 2010. Resolution of the United Nations General Assembly, A/RES/64/255, 2010 (http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/64/255 accessed 14.07.2011.
Yamamoto, T., Hashijib, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis and Prevention 40 (4), 1320–1329.
Ye, F., Lord, D. Investigating the effects of underreporting of crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit and mixed logit models. Transportation Research Record (Forthcoming). Paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C.