Accident Analysis and Prevention 47 (2012) 36–44

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Analysis of driver injury severity in rural single-vehicle crashes

Yuanchang Xie[a,*], Kaiguang Zhao[b], Nathan Huynh[c]

[a] Department of Civil and Environmental Engineering, University of Massachusetts Lowell, Lowell, MA 01854, United States
[b] Nicholas School of the Environment, Duke University, Durham, NC 27708, United States
[c] Department of Civil and Environmental Engineering, University of South Carolina, Columbia, SC 29208, United States

ARTICLE INFO

Article history:
Received 2 June 2011
Received in revised form
18 December 2011
Accepted 31 December 2011

Keywords:
Latent class logit model
Multinomial logit model
Injury severity
Rural traffic safety

ABSTRACT

Rural roads carry less than fifty percent of the traffic in the United States. However, more than half of the traffic accident fatalities occurred on rural roads. This research focuses on analyzing injury severities involving single-vehicle crashes on rural roads, utilizing a latent class logit (LCL) model. Similar to multinomial logit (MNL) models, the LCL model has the advantage of not restricting the coefficients of each explanatory variable in different severity functions to be the same, making it possible to identify the impacts of the same explanatory variable on different injury outcomes. In addition, its unique model structure allows the LCL model to better address issues pertinent to the independence from irrelevant alternatives (IIA) property. A MNL model is also included as the benchmark simply because of its popularity in injury severity modeling. The model fitting results of the MNL and LCL models are presented and discussed. Key injury severity impact factors are identified for rural single-vehicle crashes. Also, a comparison of the model fitting, analysis marginal effects, and prediction performance of the MNL and LCL models are conducted, suggesting that the LCL model may be another viable modeling alternative for crash-severity analysis.

Published by Elsevier Ltd.

1. Introduction

According to the US Department of Transportation Rural Safety Initiative released in February of 2008 (USDOT, 2008), "*Rural roads carry less than half of America's traffic yet they account for over half of the nation's vehicular deaths.*" From 1997 to 2006, traffic fatality rates on rural roads have consistently been more than twice of those on urban roads in the United States. Among all the states, the 5-year average data from 2002 to 2006 show that Florida has the nation's highest rural traffic fatality rate, which is 3.54 fatalities per 100 million vehicle miles traveled (VMT), while the national average rate during this period is 2.32 fatalities per 100 million VMT.

Table 1 shows a comparison of urban and rural fatal crash rates in Florida in 2005. It can be seen that more than 60% of the fatal crashes occurred on rural roads. Moreover, for every 100 crashes on rural roads, about two of them are fatal crashes; while for every 100 crashes on urban roads, only one of them is a fatal crash. The significant difference between rural and urban roads' crash-fatality ratio necessitates the need to better understand the underlying reasons for the higher fatality rates on rural roads and to investigate their injury severity characteristics.

Many statistical methods have been applied to traffic crash injury severity modeling, including ordered probit model (Abdel-Aty, 2003; Xie et al., 2009), ordered logit model (O'Donnell and Connor, 1996), multinomial logit model (MNL) (Savolainen and Mannering, 2007; Khorashadi et al., 2005), nested logit model (NL) (Shankar et al., 1996), ordered mixed logit model (Srinivasan, 2002), heteroscedastic ordered logit model (Wang and Kockelman, 2005), and logistics regression (Al-Ghamdi, 2002). A more comprehensive review of crash injury severity models can found in (Savolainen et al., 2011). Among these models, the ordered logit/probit models and the MNL are the most widely used ones. In the ordered logit/probit models, each explanatory variable has one coefficient, which means that the effects of this particular variable on all injury outcomes are restricted to be the same. In the MNL model, each injury outcome has a separate severity function (i.e., utility function in discrete choice modeling literature) and two severity functions can include different sets of explanatory variables. This modeling structure is quite flexible and can readily handle the distinct effects of the same variable on different injury outcomes. Although the MNL model has some advantages in terms of flexible model structure, it has certain limitations due to its independence from irrelevant alternatives (IIA) property, which originates from the independence and identical distribution (IID) assumption of the error terms in each severity function. This limitation of the MNL model is demonstrated in a previous study by Abdel-Aty (2003). In his research, Abdel-Aty compared ordered probit, MNL, and nested logit models for injury severity analysis. His research finding

* Corresponding author. Tel.: +1 9789343681; fax: +1 9789343052.
E-mail addresses: yuanchang_xie@uml.edu (Y. Xie), kz22@duke.edu (K. Zhao), huynhn@cec.sc.edu (N. Huynh).

0001-4575/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.aap.2011.12.012

Table 1
Comparison of crash fatality ratio in Florida in 2005.

	Rural	Urban
Total # of crashes (A)	117,721	150,804
# of fatal crashes (B)	2014	1170
% of fatal crashes ($B \times 100/A$)	1.71	0.78

suggested that the MNL model produced even worse fitting results than the ordered probit model. Although the nested logit model generated slightly better fitting results than the ordered probit model, the author still recommended the ordered probit model for their study after considering the difficulty in specifying the nested structure. In this study, a latent class logit (LCL) model is introduced to model traffic crash injury severity data. The LCL model is based on the MNL model. Similar to the standard MNL model, the LCL model has a flexible structure that can readily take into account the different effects of the same variable on each injury outcome. The key benefit of the LCL over the MNL is that its special structure has the potential to overcome the problems associated with the IIA property.

The rest of this paper is organized as follows. In Section 2, the MNL model is briefly discussed and the LCL model is introduced. The crash data set used in this study is then described in Section 3. In Section 4, both the MNL model and the LCL model are applied to the traffic crash injury data set. We also conduct marginal effects and prediction accuracy analyses and the results are presented in Section 5. Section 6 summarizes the major findings of this research.

2. Methodological background

2.1. Multinomial logit model

To better explain the proposed LCL model and also to make this paper self-contained, a very brief description of the standard MNL model is provided here. More details about the MNL model and the fundamental theory behind it can be found in (Train, 2003). For each single-vehicle traffic crash, assume there are k possible injury outcomes for the driver. The MNL model first constructs a severity function for each injury outcome as in Eq. (1).

$$U_{ij} = V_{ij}(\beta) + \varepsilon_{ij} \tag{1}$$

where U_{ij} is the severity function for the jth possible injury outcome of the ith driver involved in a traffic crash, with $i = 1,\ldots, n$ and $j = 1,\ldots, k$; $V_{ij}(\beta)$ is a linear-in-parameters combination of explanatory variables and is the deterministic part of the severity; β is a coefficient vector; ε_{ij} is an independent and identically distributed random variable following Gumbel distribution. Given the estimated coefficient vector β, the probability that the jth injury outcome may happen is:

$$\begin{aligned}\text{Prob}(j|\beta) &= \text{Prob}(V_{ij}(\beta) + \varepsilon_{ij} > V_{it}(\beta) + \varepsilon_{it},\\ &\forall t \neq j|\beta) = \frac{\exp(V_{ij}(\beta))}{\sum_{m=1}^{k}\exp(V_{im}(\beta))}\end{aligned} \tag{2}$$

One important assumption of the MNL model is that the random terms, ε_{ij}, of each severity functions are independent and identically distributed (IID). However, this often is not the case due to many possible reasons. For instance, traffic crash injury severity is affected by various contributing factors. Therefore, the deterministic parts of each severity function, $V_{ij}(\beta)$, should consist of many explanatory variables. However, in real world applications, it is very difficult to identify and collect all the relevant input data and include them in the severity functions. If some important explanatory variables are not included, the unobserved random portions of these severity functions are likely to be correlated, which leads to the violation of the fundamental IID assumption. The violation of the IIA property or IID assumption may lead to biased parameter estimates. It can also generate systematic errors in the choice probabilities, and a typical example is the famous red-bus-blue-bus problem. When the violation of IID assumption happens, one can choose to use a different type of model that is able to handle the correlation among the random terms of different alternatives. Another option is to modify the deterministic portions of the severity functions to capture the unobserved correlation, so that the remaining random terms can become independent. To address the potential IID assumption violation in this research, the LCL model is proposed for modeling traffic crash injury severity.

2.2. Latent class model

The LCL model can be considered as a special form of the mixed MNL model. For a typical mixed MNL model, the probability that injury outcome j will happen is described in Eq. (3).

$$\text{Prob}(j) = \int \text{Prob}(j|\beta)f(\beta)d\beta \tag{3}$$

A major difference between the standard MNL model and the mixed MNL model is the coefficient vector β. The standard MNL model assumes a constant β vector, while the mixed MNL model considers vector β as a mixture of random coefficients (φ) and constants (α). By so doing, the initial severity function becomes:

$$U_{ij} = V_{ij}(\beta) + \varepsilon_{ij} = \alpha^T W_{ij} + \varphi^T X_{ij} + \varepsilon_{ij} \tag{4}$$

where X_{ij} is a set of explanatory variables with random parameters and W_{ij} represents the explanatory variables with fixed parameters. By including the random coefficients, different injury severity outcomes become correlated even though their error terms, ε_{ij}, are still assumed to be independent and identically distributed. This is because $cov(U_{ij}, U_{ik}) = E(\varphi^T X_{ij} + \varepsilon_{ij})(\varphi^T X_{ik} + \varepsilon_{ik}) = X_{ij}^T \Omega X_{ik}$ (Train, 2003). Such a correlation can be very useful in addressing the aforementioned IID/IIA problem.

The mixed MNL model was first introduced into transportation research in 1980 (Boyd and Mellman, 1980; Cardell and Dunbar, 1980). It has since been applied to a number of areas due to the wide availability of computer simulation (Train, 2003; Gkritza and Mannering, 2008; Milton et al., 2008; Bhat, 1998; Brownstone and Train, 1998; McFadden and Train, 2000). To apply the mixed MNL model, distributions of each random coefficient in vector β must be explicitly specified, which is not a trivial task. To get around this problem, the LCL model was proposed (Swait, 1994; Greene and Hensher, 2003), which can be considered as a special form of the mixed MNL model. In the LCL model, β takes a finite set of values and the integral in Eq. (3) is replaced by a summation of weighted $\text{Prob}(j|\beta)$ over all β values. In this case, the probability for injury outcome j to happen is

$$\text{Prob}(j) = \sum_{m=1}^{M}\text{Prob}(\text{class} = m) \times \text{Prob}(j|\beta_m) \tag{5}$$

The LCL model assumes that the entire crash data set can be categorized into M different classes. Each crash event belongs to different classes with certain probabilities that are not revealed to the analyst. $\text{Prob}(j|\beta_m)$ in Eq. (5) can be determined similarly as $\text{Prob}(j|\beta)$ in Eq. (2) with β replaced by β_m. $\text{Prob}(\text{class} = m)$ is the probability that a crash event belongs to class m and can be determined by Eq. (6).

$$\text{Prob}(\text{class} = m) = \frac{\exp(V_{im}(\theta))}{\sum_{c=1}^{M}\exp(V_{ic}(\theta))} \tag{6}$$

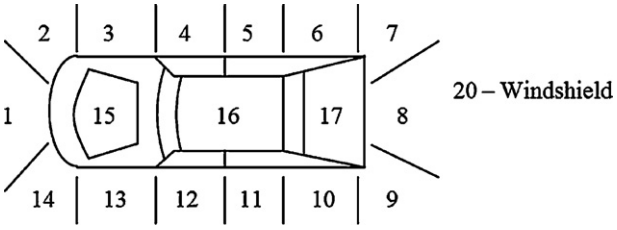

Fig. 1. Definition of points of impact.

It can be seen that the class probability Prob(class = m) is also determined based on the MNL framework. $V_{im}(\theta)$ in Eq. (6) can be a linear-in-parameters combination of a constant and several covariates. In case that no appropriate covariates can be identified to enter $V_{im}(\theta)$, only a constant needs to be chosen. Compared to the mixed MNL model, the LCL model takes only a finite set of parameters β. This can potentially save the computation time for model fitting. In addition, the LCL model can avoid the trouble of specifying the probability distributions of each random coefficient. For more details about LCL models, readers may refer to (Swait, 1994; Greene and Hensher, 2003). In this research, the LCL model will be applied to a traffic crash injury severity data set collected in Florida. The MNL model will also be applied to the same data. Our focus is not to compare these two models and prove which one is better. Including the MNL model is simply to give readers a better idea of how the LCL model performs relative to one of the most popular crash injury severity models.

3. Data description

As mentioned in Section 1, the focus of our analysis is on driver injury severities involving single-vehicle crashes on rural roads. The analysis uses the injury severity data from Florida given that it has the nation's highest rural traffic crash fatality rate. Since different types of vehicles and crash locations may exhibit quite different injury level characteristics, we further narrowed down the scope of this study to automobile and van related crashes. Also, only crashes occurred on interstates, US and state highways are considered. The total number of crashes with valid data is 4285, and more than 50 explanatory variables considered are listed and defined in Table 2. More information about the definitions of these variables can be found in the Florida Traffic Crash Records Database Codebook (OMRD, 2006). As done in many previous studies, five crash injury outcomes are considered in this research: "no injury", "possible injury", "non-incapacitated injury", "incapacitated injury", and "fatal injury".

4. Model fitting results

4.1. MNL model estimation result

Before presenting the MNL results, it is important to note that we conducted the Small–Hsiao test (Small and Hsiao, 1985) and t test for the logsum parameters of several nested logit models to determine if an independence of irrelevant alternatives (IIA) violation exists in the MNL formulations. The results show that the IIA property of the MNL model is violated. Nevertheless, we present the MNL estimation results as a basis of comparison for the LCL model.

When fitting the MNL model, only variables available in the 2005 Florida crash data set are considered. The Florida traffic crash records database contains many detailed information. In this research, those explanatory variables that could potentially affect injury level are used for fitting the MNL model and are listed in Table 2. For the five crash injury outcomes (alternatives), the non-incapacitated injury outcome is used as the base alternative. Since the values of all explanatory variables do not vary across alternatives (i.e., injury outcomes), each variable in the MNL model can have up to four coefficients associated with it. Coefficients that are statistically significant at the 0.05 level are selected and included in the model shown in Table 3. As can be seen, 31 out of the 53 explanatory variables are finally selected.

Most estimated coefficients in Table 3 have the expected signs. For example, the result for *DAge_1* and *DAge_2* suggests that young drivers (*DAge_1*) involved in crashes tend to receive less severe injuries and senior drivers (*DAge_2*) are likely to be involved in more severe crashes, which is logical. The negative coefficients for *White*, *Black*, and *Hispanic* suggest that drivers in these ethnic groups are less prone to fatal injuries. Compared to white and Hispanic drivers, black drivers have a negligibly higher chance to be involved in fatal crashes. This finding is different than the results reported in some other studies (Zhu and Srinivasan, 2011; Crompton et al., 2010), which found that African–American drivers and motorcyclists are much more likely to be involved in fatal accidents than their White and Hispanic counterparts. Male drivers are found to have a higher possibility to be involved in no-injury crashes. The significant coefficients for *DUI* suggest that DUI drivers are more likely to be involved in both *incapacitated*/fatal and no-injury crashes. This finding also demonstrates the benefit of using the MNL model instead of ordered logit/probit models. The results for *Seatbelt* clearly show that wearing seat belts is critical for mitigating driver injury severity; similarly, the coefficients for *Ejected* suggest that it is very important to restrain drivers in their vehicles when crashes happen. The coefficients for *Airbag* suggest that if the airbag was deployed, the chance for the driver to be involved in no-injury or property damage only injury was smaller. This is reasonable as airbags are often deployed in more severe traffic crashes. The results for *Auto* suggest that automobile drivers usually sustain less severe injuries compared to van drivers. If the crashed vehicles remain on the roadway (*OnRoadway*), then drivers are less likely to be involved in no-injury or fatal crashes. For crashes that occur on medians, the drivers are less likely to suffer from incapacitated injury. Crashes that occur in work zones (*Workzone*) will increase drivers' chance for incapacitated and fatal injuries. For points of impact in *Area1* and *Area3*, they can all significantly increase drivers' risk of more severe injuries. Among the listed 1st harmful events (*Tree_Pole_1*, *Ditch_1*, *Overturned_1*, *Guardrail_1*, *CBarrier_1*, and *Water_1*), the most dangerous ones are collisions with tree/shrubbery/utility pole/light pole, vehicle overturn, collisions with concrete barrier, and run into water or ditch. Among the 2nd harmful events, collision with tree/shrubbery/utility pole/light pole, run into ditch, and vehicle overturn are found to substantially decrease drivers' possibilities of less severe injuries. It is interesting that darkness can in fact increase drivers' probabilities of being involved in no-injury crashes. This is probably because drivers tend to be more cautious when it is dark without street light. As expected, level horizontal alignments and divided roadways can all decrease the chance for more severe injury. For crashes on Interstate and US highways rather than state highways, they have a slightly higher chance to be no-injury crashes. This may be explained by the higher design standard of Interstate and US highways (e.g., wider shoulders). In addition, the results suggest that crashes on Interstate highways are more likely to be fatal. This could be caused by the higher average speed on Interstate highways. A counter-intuitive finding is that higher posted speed limit may slightly (the corresponding coefficient is 0.009) increase the chance for no-injury crashes. Intuitively, this is not plausible. A possible explanation is that the variables *Post_mph*, *Interstate*, and *US_Highway* are correlated (i.e., higher posted speed limits for Interstate and US highways), and the positive effects of *Interstate* and *US_Highway* as discussed earlier are erroneously picked up by

Table 2
Explanatory variables considered in this study.

Variable type	Variable name	Description
Driver information	*DAge_1*	If driver age is ≤24 (1 = Yes, 0 = No)
	DAge_2	If driver age is ≥65 (1 = Yes, 0 = No)
	White	If driver is white (1 = Yes, 0 = No)
	Black	If driver is black (1 = Yes, 0 = No)
	Hispanic	If driver is Hispanic (1 = Yes, 0 = No)
	Male	If driver is male (1 = Yes, 0 = No)
	DUI	If driver was under the influence of alcohol or drugs (1 = Yes, 0 = No)
	Fatigue	If driver was fatigue or asleep (1 = Yes, 0 = No)
	Seatbelt	If seatbelt was used (1 = Yes, 0 = No)
Driver information	*Airbag*	If airbag was deployed (1 = Yes, 0 = No)
	Ejected	If driver was either completed or partially ejected from the vehicle (1 = Yes, 0 = No)
Vehicle information	*Veh_Age*	Vehicle age
	Auto	1 = It was an automobile, 0 = it was a van
	Tire_Defect	1 = Vehicle had worn tire(s), 0 = otherwise
Crash information	*OnRoadway*	If crash occurred on roadway (1 = Yes, 0 = No)
	Shoulder	If crash occurred on shoulder (1 = Yes, 0 = No)
	Median	If crash occurred on median (1 = Yes, 0 = No)
	Turnlane	If crash occurred on turnlane (1 = Yes, 0 = No)
	Violation	1 = If at least one moving violation cited in crash, =otherwise
	Workzone	If crash occurred in a workzone (1 = Yes, 0 = No)
	Area1	1 = Points of impact included 11 and 12 (see Fig. 1 for detailed definition of points of impact), 0 = otherwise
	Area2	1 = Points of impact included 4 and 5, 0 = otherwise
	Area3	1 = Points of impact included 13 and 15, 0 = otherwise
	Area4	1 = Points of impact included 1, 2, 3, and 14, 0 = otherwise
	Area5	1 = Points of impact included 16 and 20, 0 = otherwise
	Animal_1	1 = If 1st harmful event was collision with animal, 0 = otherwise
	Tree_Pole_1	1 = If 1st harmful event was collision with tree/shrubbery/utility pole/light pole, 0 = otherwise
	Ditch_1	1 = If 1st harmful event was running into ditch/culvert, 0 = otherwise
	Overturned_1	1 = If 1st harmful event was overturned, 0 = otherwise
	Guardrail_1	1 = If 1st harmful event was collision with guardrail, 0 = otherwise
	CBarrier_1	1 = If 1st harmful event was collision with concrete barrier wall, 0 = otherwise
	Water_1	1 = If 1st harmful event was running off road into water, 0 = otherwise
	Tree_2	1 = If 2nd harmful event was collision with tree/shrubbery, 0 = otherwise
	Ditch_2	1 = If 2nd harmful event was running into ditch/culvert, 0 = otherwise
	Overturned_2	1 = If 2nd harmful event was overturned, 0 = otherwise
Weather and lighting	*DayLight*	1 = If lighting was day light, 0 = otherwise
	Dark	1 = If it was dark without street lighting, 0 = otherwise
	Rain	1 = If it was raining, 0 = otherwise
	Fog	1 = If it was fogging, 0 = otherwise
Roadway information	*Pave_Shoulder*	1 = If shoulder was paved, 0 = otherwise
	Dry_Surf	1 = If roadway surface was dry, 0 = otherwise
	Surf_Defect	1 = If roadway surface had defects, 0 = otherwise
	Curve_Level	1 = If crash was on a level curve, 0 = otherwise
	Curve_Grade	1 = If crash was on a curve with grade, 0 = otherwise
	Curve	1 = If crash was on a curve, 0 = otherwise
	Level	1 = If crash was on a level segment, 0 = otherwise
	Concrete	1 = If roadway had concrete surface, 0 = otherwise
	Divided	1 = If two travel directions were divided, 0 = otherwise
	Interstate	1 = If crash occurred on interstate highway, 0 = otherwise
	US_Highway	1 = If crash occurred on US highway, 0 = otherwise
	Intersection	1 = If crash occurred at intersection, 0 = otherwise
Speed information	*Post_mph*	Posted speed limit value
	Est_mph	Estimated vehicle speed prior to crash

Post_mph. Intuitively the higher the vehicle speed prior to crash is, the more severe the crash will be. This assumption has been supported by the coefficients for *Est_mph*, although the magnitudes of the coefficients suggest that this impact is almost negligible. One possible reason is that drivers involved in crashes tend to underreport their traveling speeds, and this may affect the speed recorded by the police officer in the accident report.

Quite a few explanatory variables are found to be insignificant at the 0.05 level in explaining crash injury severity: *Veh_Age*, *Fatigue*, *Tire_Defect*, *Shoulder*, *Turnlane*, *Violation*, *Area2*, *Area4*, *Area5*, *Animal_1*, *Guardrail_1*, *DayLight*, *Rain*, *Fog*, *Pave_Shoulder*, *Dry_Surf*, *Surf_Defect*, *Curve_Level*, *Curve_Grade*, *Curve*, *Concrete*, and *Intersection*. It is interesting that *Fog* is among the insignificant variables. One potential reason is the small sample size. Only 61 out of the

Table 3
MNL model estimation result.

Variable name	No injury		Possible injury		Incap. injury		Fatal injury	
	coef	*b*/std	coef	*b*/std	coef	*b*/std	coef	*b*/std
DAge_1	–	–	–	–	−0.343	−3.479	−0.846	−3.336
DAge_2	−0.424	−2.759	–	–	–	–	1.685	4.706
White	–	–	–	–	–	–	−2.452	−7.962
Black	–	–	–	–	–	–	−1.900	−5.491
Hispanic	–	–	–	–	–	–	−2.523	−5.885
Male	0.576	7.988	–	–	–	–	–	–
DUI	1.212	8.018	–	–	0.606	3.322	2.619	10.011
Seatbelt	1.448	10.635	0.753	6.254	−0.678	−6.240	−1.801	−7.279
Airbag	−1.138	−11.318	−0.500	−5.124	–	–	–	–
Ejected	−1.103	−2.385	−0.867	−2.150	1.003	4.474	2.481	7.982
Auto	–	–	–	–	−0.354	−3.200	–	–
OnRoadway	−0.241	−2.874	–	–	–	–	−0.565	−2.211
Median	–	–	–	–	−0.456	−3.078	–	–
Workzone	−0.339	−2.157	−0.592	−3.430	–	–	–	–
Area1	–	–	–	–	0.568	3.014	0.995	2.580
Area3	–	–	–	–	–	–	1.179	2.571
Tree_Pole_1	−0.597	−4.994	−0.379	−3.184	0.547	3.981	1.211	4.522
Ditch_1	–	–	–	–	0.525	3.991	–	–
Overturned_1	−1.507	−11.622	−0.560	−4.557	0.464	3.409	–	–
CBarrier_1	−0.475	−3.359	–	–	−0.585	−2.320	–	–
Water_1	–	–	–	–	–	–	1.167	1.982
Tree_2	−0.712	−4.360	−0.440	−2.771	–	–	–	–
Ditch_2	−0.426	−2.113	–	–	–	–	–	–
Overturned_2	−1.576	−11.264	−0.662	−5.297	–	–	–	–
Dark	0.243	3.054	–	–	–	–	–	–
Level	0.288	3.143	–	–	–	–	–	–
Divided	0.389	4.237	–	–	–	–	–	–
Interstate	0.363	3.862	–	–	–	–	0.514	2.332
US_Highway	0.205	1.922	–	–	–	–	–	–
Post_mph	0.009	2.058	–	–	–	–	–	–
Est_mph	−0.028	−7.333	−0.005	−2.357	–	–	–	–
Number of observations							4285	
Log-likelihood at zero							−6896.441	
Log-likelihood at convergence							−5328.343	
Likelihood ratio index ρ (McFadden Pseudo R-Square)							0.227	

4285 samples are fog related. Another possible explanation is that fog may significantly increase the number of traffic crashes. However, it may not necessarily increase the injury severity level, since drivers tend to be more cautious when driving in fog.

4.2. Latent class logit (LCL) model estimation result

The LCL model is fitted with exactly the same set of explanatory variables as the MNL model described above. Tables 4 and 5 show the estimation results for the LCL model. As discussed in the methodological background, the latent classification of each crash is not revealed to the analyst. However, the analyst can pre-specify the number of classes. There are no rigorous rules regarding how to select the number of classes. In general, the more classes the more complicated the model would be. In this study, two classes are considered. No explanatory variables are selected to enter $V_{im}(\theta)$ (see Eq. (6)), and only a constant is used in each class severity function. Therefore, for all crashes considered, their estimated probabilities to be in each class are the same. These probabilities are shown in Table 5.

The likelihood ratio index ρ (also referred to as McFadden Pseudo R-Square) in Table 4 is 0.237, which is slightly better than the likelihood ratio index for the MNL model, which is encouraging. However, we have to be cautious and avoid making any definitive conclusions as this improvement may be data-specific. Different from the results in Table 3, each variable in Table 4 has two sets of coefficients associated with it. The magnitudes and signs of these two sets of coefficients can be quite different, which makes model interpretation a bit more challenging. For instance, consider the negative signs for the coefficients of variable *Water_1* (class 2 and "fatal injury" outcome). This seems to suggest that if the 1st harmful event is running into water, the chance for "fatal injury" would be reduced, which is counter-intuitive. This is because the prediction from the LCL model is a weighted average of probabilities from all classes. The impact of an explanatory variable cannot be determined simply based on the results of a single class. Thus, one should also look at the coefficients of *Water_1* for class 1 as well. In addition, the interpretation of the coefficients from MNL and LCL models is different from ordered logit/probit models. For ordered logit/probit models, there is only one severity function, and a negative coefficient simply increases the probability for less severe injury outcomes (i.e., "no injury") (Washington et al., 2003). While for MNL and LCL models, the effect of an explanatory variable needs to be determined by taking into account corresponding coefficients in all severity functions.

The example illustrated in the previous paragraph shows that although the MNL and LCL models provide better modeling flexibility, interpreting their modeling outputs becomes less straightforward. This is particularly true for the LCL model, especially when several classes are specified. To address this problem, marginal effect analysis is introduced, which is often used by traffic safety researchers (Ulfarsson and Mannering, 2004). In the next section, marginal effects are calculated for selected explanatory variables to demonstrate how they can be used to interpret the results of MNL and LCL models.

Table 4
LCL model estimation result.

Variable name	No injury		Possible Injury		Incap. injury		Fatal injury	
	coef	*b*/std	coef	*b*/std	coef	*b*/std	coef	*b*/std
Severity parameters in latent class 1								
DAge_1	–	–	–	–	−0.357	−2.584	−3.649	−3.103
DAge_2	−0.034	−0.109	–	–	–	–	0.938	0.677
White	–	–	–	–	–	–	−3.367	−4.410
Black	–	–	–	–	–	–	−2.538	−3.269
Hispanic	–	–	–	–	–	–	−6.201	−4.058
Male	0.430	2.911	–	–	–	–	–	–
DUI	2.430	6.329	–	–	0.485	1.572	5.111	6.150
Seatbelt	2.503	8.755	0.650	3.618	−0.358	−2.093	−0.332	−0.614
Airbag	−2.227	−9.626	−0.670	−4.182	–	–	–	–
Ejected	−2.348	−2.757	−6.778	−0.542	0.792	2.715	4.170	4.828
Auto	–	–	–	–	−0.201	−1.186	–	–
OnRoadway	−0.720	−4.042	–	–	–	–	−2.417	−3.444
Median	–	–	–	–	−0.804	−3.222	–	–
Workzone	−0.290	−0.905	0.061	0.264	–	–	–	–
Area1	–	–	–	–	0.373	1.451	0.980	1.213
Area3	–	–	–	–	–	–	−2.227	−0.456
Tree_Pole_1	−1.757	−6.377	−0.538	−2.625	0.600	3.132	0.022	0.029
Ditch_1	–	–	–	–	−0.031	−0.129	–	–
Overturned_1	−5.161	−5.591	−0.529	−2.741	0.220	1.093	–	–
CBarrier_1	−1.013	−3.880	–	–	−2.139	−1.668	–	–
Water_1	–	–	–	–	–	–	3.265	3.438
Tree_2	−3.296	−5.390	−0.476	−2.304	–	–	–	–
Ditch_2	−0.428	−1.154	–	–	–	–	–	–
Overturned_2	−3.108	−8.698	−0.358	−1.830	–	–	–	–
Dark	0.372	2.168	–	–	–	–	–	–
Level	−0.105	−0.586	–	–	–	–	–	–
Divided	0.098	0.519	–	–	–	–	–	–
Interstate	0.875	4.221	–	–	–	–	0.480	0.861
US_Highway	0.070	0.321	–	–	–	–	–	–
Post_mph	−0.017	−1.918	–	–	–	–	–	–
Est_mph	0.012	1.529	−0.001	−0.267	–	–	–	–
Severity parameters in latent class 2								
DAge_1	–	–	–	–	−0.312	−1.817	−0.186	−0.695
DAge_2	−1.930	−5.782	–	–	–	–	2.236	5.416
White	–	–	–	–	–	–	−2.797	−6.705
Black	–	–	–	–	–	–	−2.212	−4.792
Hispanic	–	–	–	–	–	–	−2.324	−4.856
Male	1.296	9.832	–	–	–	–	–	–
DUI	0.425	1.713	–	–	0.785	3.320	1.851	5.761
Seatbelt	0.590	2.870	1.323	5.803	−1.348	−7.285	−2.859	−9.532
Airbag	0.101	0.611	−0.349	−2.272	–	–	–	–
Ejected	0.627	0.624	3.868	4.061	3.042	3.348	4.695	5.066
Auto	–	–	–	–	−0.565	−3.087	–	–
OnRoadway	0.284	1.837	–	–	–	–	0.649	1.831
Median	–	–	–	–	0.347	1.559	–	–
Workzone	−0.590	−2.290	−5.184	−3.157	–	–	–	–
Area1	–	–	–	–	1.282	3.954	1.497	3.553
Area3	–	–	–	–	–	–	1.557	2.900
Tree_Pole_1	1.078	5.405	−0.210	−1.120	−2.364	−3.002	1.799	5.630
Ditch_1	–	–	–	–	1.393	6.978	–	–
Overturned_1	1.870	8.758	−1.526	−5.361	0.699	2.825	–	–
CBarrier_1	0.263	1.031	–	–	0.686	2.333	–	–
Water_1	–	–	–	–	–	–	−4.162	−0.534
Tree_2	2.574	9.210	−1.180	−2.757	–	–	–	–
Ditch_2	−0.961	−2.229	–	–	–	–	–	–
Overturned_2	0.139	0.695	−2.086	−6.955	–	–	–	–
Dark	0.293	2.112	–	–	–	–	–	–
Level	1.327	7.338	–	–	–	–	–	–
Divided	1.106	6.840	–	–	–	–	–	–
Interstate	0.108	0.670	–	–	–	–	0.578	2.333
US_Highway	0.551	3.086	–	–	–	–	–	–
Post_mph	0.067	7.120	–	–	–	–	–	–
Est_mph	−0.141	−14.398	−0.014	−3.759	–	–	–	–
Number of observations							4285	
Log-likelihood at zero							−6896.441	
Log-likelihood at convergence							−5263.962	
Likelihood ratio index ρ (McFadden Pseudo R-Square)							0.237	

Table 5
Estimated latent class probabilities.

	Probability	*b*/std
Probability in class 1	0.6505	337.632
Probability in class 2	0.3495	15.157

5. Analysis of marginal effects and prediction accuracy

5.1. Marginal effects

For both the MNL and LCL models, each explanatory variable is often associated with several coefficients. To quantify these variables' overall impacts on the crash injury outcomes, marginal effect analysis is often utilized. Several variables are selected for the marginal effect analysis to demonstrate how to interpret the modeling outputs of the MNL and LCL models. In this research, the marginal effects are calculated using Limdep®. Since the original marginal effect values are very small, these original values are multiplied by 100 and presented in Tables 6 and 7. The numbers in bold font represent direct marginal effects and the numbers in regular font represent cross-marginal effects. For the MNL model, the direct and cross-marginal effects are calculated by Eqs. (7) and (8), respectively (Greene, 2008).

$$\frac{\partial P_{ij}}{\partial x_{ijk}} = \frac{\partial V_{ij}}{\partial x_{ijk}} P_{ij}(1 - P_{ij}) = \beta_{jk} P_{ij}(1 - P_{ij}) \tag{7}$$

$$\frac{\partial P_{im}}{\partial x_{ijk}} = -\frac{\partial V_{ij}}{\partial x_{ijk}} P_{im} P_{ij} = -\beta_{jk} P_{im} P_{ij} \tag{8}$$

The direct marginal effect (Eq. (7)) represents the impact of changes in variable k of outcome j (denoted by x_{ijk}) on the probability for crash i to be in injury outcome j (denoted by P_{ij}). The cross-marginal effect (Eq. (8)) describes the impact of changes in variable k of alternative j ($j \neq m$) on the probability (P_{im}) for crash i to be in outcome m. Details on how to derive Eqs. (7) and (8) can be found in (Train, 2003) and will not be duplicated here. For the LCL model, a marginal effect is calculated for each class using the same method for the MNL model. The final marginal effect of a variable is the summation of the marginal effects for each class weighted by their posterior latent class probabilities.

In the marginal effect analysis, some selected variables enter multiple severity functions; thus, changing their values will affect all corresponding severity function values. To find out their combined effects, their marginal effects with respect to different injury outcomes need to be added together as shown in Tables 6 and 7.

An interesting finding is that the magnitude of the marginal effects in Table 7 is significantly smaller than Table 6. However, their relative values and signs are similar. To better illustrate this similarity, the marginal effects for variable *Black* are standardized based on the values in column "no injury" in Tables 6 and 7. The adjusted results are shown in Table 8. As can be seen from the data, the relative impacts of variable *Black* on different injury outcomes from the MNL and LCL models are similar. Such a trend also exists for other variables. This suggests that although the MNL and LCL models generate outputs that can be different in an order of magnitude, the relative difference between the marginal effects within each model is approximately the same.

5.2. Prediction accuracy

The previous subsection shows the discrepancies as well as consistency in the marginal effects generated by the MNL and LCL models. To further assess the performance of the LCL model, a prediction experiment is conducted to evaluate the goodness-of-fits of the two models. The original data set is separated into two groups, one for model fitting and the other for model evaluation. If a model can generate more correct predictions on both the fitting and the evaluation data sets, then it is reasonable to consider that it fits the data better.

Table 6
Analysis of marginal effects for MNL model.

Variables	In severity function of	Effects on probabilities of the following injury outcomes				
		No injury	Possible injury	Non-incap. injury	Incap. injury	Fatal injury
	Incap. injury	−1.8163	−1.6341	−2.3109	6.1871	−0.4259
Area1	Fatal injury	−0.4146	−0.2873	−0.5412	−0.7458	1.9889
	Combined effects	−2.2309	−1.9214	−2.8521	5.4413	1.5630
Ditch_1	Incap. injury	−1.6775	−1.5092	−2.1343	5.7142	−0.3933
Black	Fatal injury	0.7919	0.5487	1.0337	1.4246	−3.7989
Water_1	Fatal injury	−0.4863	−0.3369	−0.6348	−0.8748	2.3327

Table 7
Analysis of marginal effects for LCL model.

Variables	In severity function of	Effects on probabilities of the following injury outcomes				
		No injury	Possible injury	Non-incap. injury	Incap. injury	Fatal injury
	Incap. injury	−0.1032	−0.1027	−0.1670	0.4246	−0.0517
Area1	Fatal injury	−0.0433	−0.0298	−0.0533	−0.0780	0.2043
	Combined effects	−0.1465	−0.1325	−0.2203	0.3466	0.1526
Ditch_1	Incap. injury	−0.3058	−0.2849	−0.4856	1.2163	−0.1400
Black	Fatal injury	0.1612	0.1240	0.2302	0.2441	−0.7595
Water_1	Fatal injury	−0.0730	−0.0288	−0.0381	−0.0336	0.1734

Table 8
Standardized marginal effects for variable *Black*.

Model	Effects on probabilities of the following injury outcomes				
	No injury	Possible injury	Non-incap. Injury	Incap. injury	Fatal injury
MNL	1.0000	0.6929	1.3053	1.7990	−4.7972
LCL	1.0000	0.7692	1.4280	1.5143	−4.7115

Table 9
Prediction accuracies of MNL and LCL models.

	Injury severity	MNL model		LCL model	
		Mean %	Std	Mean %	Std
Model fitting	No injury	76.1	1.0	89.1	1.9
	Possible injury	12.3	4.1	51.9	8.7
	Non-incap. injury	44.5	3.7	66.0	5.4
	Incap. injury	20.2	0.7	33.8	5.7
	Fatal injury	38.7	2.7	59.1	6.0
	Overall	44.4	0.6	66.0	2.7
Prediction	No injury	74.9	2.6	86.2	2.2
	Possible injury	11.2	4.4	48.4	8.9
	Non-incap. injury	44.8	5.7	64.4	5.0
	Incap. injury	20.2	1.4	30.5	6.5
	fatal injury	30.5	3.0	44.5	6.0
	overall	43.7	1.2	63.1	2.7

From the collected data, 3000 observations are randomly drawn for model fitting, and the remaining data are used for evaluation. This process is repeated 10 times and the prediction results are shown in Table 9. Clearly, the LCL model generates very satisfying fitting and prediction results. Compared to the MNL model, the LCL model improves the prediction accuracy for the possible injury category by around 37%. For other injury outcomes, the improvements from the LCL model range between 10% and 20%, which are quite significant considering that this is the average result based on 10 randomly generated samples.

6. Summary and conclusions

This research is focused on rural single-vehicle traffic crash injury severity analysis, utilizing both the multinomial logit (MNL) model and the latent class logit (LCL) model. The contribution of this research is twofold. First, the LCL model is introduced for traffic crash injury severity analysis. Traditionally, MNL is typically used for traffic crash injury severity studies. However, its applicability is limited by the IID assumption, which assumes that all error terms in severity functions are independent and identically distributed. This IID assumption may not be valid in some traffic safety studies. LCL model has the capability to better handle the correlations among those error terms. In addition, the LCL model is easy to apply because it does not require specifying any distributions for each severity parameter.

The second major contribution of this study is to apply the LCL and MNL models to investigate rural single-vehicle crash driver injury severity. We apply both models to the 2005 Florida traffic collision data to find out the relationship between injury severities and related traffic, geometry, driver, vehicle, and surrounding environment characteristics. A total of 53 potential explanatory variables are examined initially using the MNL model; 31 of them are found to be statistically significant at the 0.05 level. Variables such as driver age, DUI, seat belt usage, points of impact, lighting condition, speed, 1st and 2nd harmful events, and ethnicity are found to be closely related to driver injury severity levels. Some commonly used variables such as vehicle age and surface condition (i.e., dry, wet, or icy) however are found to have no significant impact on driver injury severity. This finding could have some important implications in safety analysis and warrant additional research. In addition, we demonstrate how to interpret the LCL model results. Based on the selected 31 explanatory variables, a LCL model is fitted. Due to the large amount of coefficients and the interactions between severity function values, the results from the MNL and LCL model cannot be interpreted directly as in ordered logit/probit models. To address this problem, marginal effects are calculated and analyzed. The analysis of marginal effects suggests that the two models produce results of quite different magnitudes. For data set and severity functions considered in this study, the marginal effects from the LCL model are considerably smaller than those from the MNL model. However, further analysis shows that the marginal effects from both the LCL and MNL models are consistent in terms of signs and trends.

In this research, the prediction accuracies of the MNL and LCL models are also evaluated. The result shows that the fitting and prediction performance of LCL model is satisfactory compared to the MNL model, suggesting that the LCL model fits the data well and is a promising tool for future crash injury severity studies. If the primary purpose of a study is to predict injury severity outcomes, the LCL model can be a very good choice. Although the result is encouraging, we need to acknowledge that the MNL model is selected as the benchmark in this study mainly because of its popularity; a comparison with the widely used MNL model can give readers a better idea of the performance of the LCL model. This comparison is not intended to prove that the LCL model is the best tool for this particular data set because there might be other existing models that fit the data better than the MNL model. However, it is out of the scope of this study to conduct a comprehensive comparison of existing crash injury models and different severity functions. Interested readers may refer to Savolainen et al. (2011) and Ye and Lord (2011) for more information on model comparison. Also, the model fitting and prediction comparison result should not be used to draw a general conclusion that the LCL model has better prediction performance than the MNL model, even though the comparison in this study is based on 10 randomly generated samples. Additional tests on other data sets should be conducted before such a conclusion can be drawn. For future research, an interesting direction would be to explore some other methods such as the mixed logit model and to compare them with the LCL model using representative data sets.

Acknowledgments

The first author would like to thank the Florida Department of Highway Safety and Motor Vehicles for providing the data used in this research. Also, the authors would like to thank the three anonymous reviewers for their constructive comments. The results and opinions expressed in this paper are solely of the authors.

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered Probit models. J. Safety Res. 34 (5), 597–603.
Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accid. Anal. Prev. 34 (6), 729–741.
Bhat, C., 1998. Accommodating variations in responsiveness to level-of-service variables in travel mode choice models. Transport. Res. A 32 (7), 455–507.
Boyd, J., Mellman, J., 1980. The effect of fuel economy standards on the US automotive market: a hedonic demand analysis. Transport. Res. A 14 (5–6), 367–378.
Brownstone, D., Train, K., 1998. Forecasting new product penetration with flexible substitution patterns. J. Econometrics 89 (1–2), 109–129.

Cardell, S., Dunbar, F., 1980. Measuring the societal impacts of automobile downsizing. Transport. Res. A 14 (5–6), 423–434.
Crompton, J.G., Pollack, K.M., Oyetunji, T., Chang, D.C., Efron, D.T., Haut, E.R., Cornwell, E.E., Haider, A.H., 2010. Racial disparities in motorcycle-related mortality: an analysis of the National Trauma Data Bank. Am. J. Surg. 200 (2), 191–196.
Gkritza, K., Mannering, F.L., 2008. Mixed logit analysis of safety-belt use in single- and multi-occupant vehicles. Accid. Anal. Prev. 40 (2), 443–451.
Greene, W.H., 2008. Econometric Software, Inc. NLOGIT Version 4.0 Reference Guide.
Greene, W.H., Hensher, D.A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. Transport. Res. B 37 (8), 681–698.
Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F.L., 2005. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accid. Anal. Prev. 37 (5), 910–921.
McFadden, D., Train, K., 2000. Mixed MNL models of discrete response. J. Appl. Econometrics 15 (5), 447–470.
Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40 (1), 260–266.
O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accid. Anal. Prev. 28 (6), 739–753.
Office of Management Research and Development (OMRD), 2006. Florida Department of Highway Safety and Motor Vehicles. The Florida Traffic Crash Records Database Codebook.
Savolainen, P.T., Mannering, F.L., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accid. Anal. Prev. 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. Accid. Anal. Prev. 43 (3), 1666–1676.
Shankar, V., Mannering, F.L., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accid. Anal. Prev. 28 (3), 391–401.
Small, K., Hsiao, C., 1985. Multinomial logit specification tests. Int. Econ. Rev. 26 (3), 619–627.
Srinivasan, K.K., 2002. Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. Transport. Res. Rec. 1784, 132–141.
Swait, J., 1994. A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data. J. Retail Consum. Serv. 1 (2), 77–89.
Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, New York.
Ulfarsson, G.F., Mannering, F.L., 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accid. Anal. Prev. 36 (2), 135–147.
The US Department of Transportation Rural Safety Initiative, US DOT, 2008. Retrieved from: http://www.dot.gov/affairs/ruralsafety/ruralsafetyinitiativeplan.htm.
Wang, X.K., Kockelman, K.M., 2005. Occupant injury severity using a heteroscedastic ordered logit model: distinguishing the effects of vehicle weight and type. Transport. Res. Rec. 1908, 195–204.
Washington, S., Karlaftis, M.G., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC.
Xie, Y.C., Zhang, Y.L., Liang, F.M., 2009. Crash injury severity analysis using Bayesian ordered probit model. J. Transport. Eng. 135 (1), 18–25.
Ye, F., Lord, D., 2011. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. In: Proceedings of the 90th Annual Meeting of the Transportation Research Board, Washington, DC.
Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accid. Anal. Prev. 43 (1), 49–57.