Accident Analysis and Prevention 43 (2011) 1140–1147

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data

Panagiotis Ch. Anastasopoulos[a,b,*], Fred L. Mannering[a,b,1]

[a] School of Civil Engineering, Purdue University, West Lafayette, IN 47907-2051, United States
[b] Center for Road Safety, Purdue University, West Lafayette, IN 47907-2051, United States

ARTICLE INFO

Article history:
Received 15 August 2010
Received in revised form
13 December 2010
Accepted 18 December 2010

Keywords:
Crash-injury severities
Random parameters logit model
Mixed logit model
Pavement condition
Roadway geometrics

ABSTRACT

Traditional crash-severity modeling uses detailed data gathered after a crash has occurred (number of vehicles involved, age of occupants, weather conditions at the time of the crash, types of vehicles involved, crash type, occupant restraint use, airbag deployment, etc.) to predict the level of occupant injury. However, for prediction purposes, the use of such detailed data makes assessing the impact of alternate safety countermeasures exceedingly difficult due to the large number of variables that need to be known. Using 5-year data from interstate highways in Indiana, this study explores fixed and random parameter statistical models using detailed crash-specific data and data that include the injury outcome of the crash but not other detailed crash-specific data (only more general data are used such as roadway geometrics, pavement condition and general weather and traffic characteristics). The analysis shows that, while models that do not use detailed crash-specific data do not perform as well as those that do, random parameter models using less detailed data still can provide a reasonable level of accuracy.

1. Introduction

The most common approach to statistically modeling the frequency of crashes and their injury severity is to start with a crash-frequency model (to study the number of crashes that have occurred on a specific road segment or intersection in some time period) and then consider the injury severity of the crash (often defined is the most severe occupant injury observed in the crash) conditional on the crash having occurred. By considering this conditional crash-specific severity, the many details relating to the crash (number of vehicles involved, age of occupants, weather conditions at the time of the crash, types of vehicles involved, crash type, occupant restraint use, airbag deployment, etc.) can be considered in the development of a statistical model of severity using detailed crash reports. However, the use of such detailed data makes it difficult to forecast changes in injury severities because crash-specific data elements need to be known. Given this, there is considerable appeal in developing some combination of frequency and severity models that are less data intensive than the traditional conditional-severity approach. Such models could be readily used for network screening (ranking of sites in need of safety countermeasures) and help identify at least some influential factors that affect the level of injury-severities.

Models that look at the total frequency of all crashes (without distinguishing by injury severity) are abundant in the literature and include a wide variety of modeling approaches including: Poisson models; negative binomial models; Poisson-lognormal models; zero-inflated count models; Conway–Maxwell–Poisson models; Gamma models; generalized estimating equation models; generalized additive models; random effects models; negative multinomial models; random parameters models; and finite mixture and Markov switching models (see Lord and Mannering, 2010 for a complete review of this literature).

In terms of combining crash frequencies and severity in the modeling process, several researchers have approached the frequency of crashes while simultaneously considering their severity in an effort to study crash counts at specific severity levels. This complicates the modeling process considerably because the counts of crashes in each injury-severity level are not independent – thus the correlation among specific injury crash counts (no-injury, possible injury, evident injury, disabling injury, fatality) must be considered. A variety of multivariate methods have been used to address this problem including the multivariate Poisson, multivariate negative binomial, and the multivariate Poisson-lognormal (Miaou and Lord, 2003; Miaou and Song, 2005; Bijleveld, 2005; Song et al., 2006; Ma and Kockelman, 2006; Park and Lord, 2007; Bonneson and Pratt, 2008; Geedipally and Lord, 2010; Ma et al., 2008; Depaire et al., 2008; Ye et al., 2009; Aguero-Valverde and Jovanis, 2009; El-Basyouny and Sayed, 2009).

* Corresponding author at: School of Civil Engineering, Purdue University, West Lafayette, IN 47907-2051, United States. Tel.: +1 512 810 3226; fax: +1 765 494 0395.
E-mail addresses: panast@purdue.edu (P.Ch. Anastasopoulos), flm@purdue.edu (F.L. Mannering).
[1] Tel.: +1 765 496 7913; fax: +1 765 494 0395.

0001-4575/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2010.12.024

Traditional crash-severity models (those which are conditioned on the crash having occurred), have included a wide variety of methodological approaches such as multinomial logit models, dual-state multinomial logit models, nested logit models, mixed logit models and ordered probit models (Shankar et al., 1996; Duncan et al., 1998; Chang and Mannering, 1999; Carson and Mannering, 2001; Khattak, 2001; Khattak et al., 2002; Kockelman and Kweon, 2002; Lee and Mannering, 2002; Abdel-Aty, 2003; Kweon and Kockelman, 2003; Ulfarsson and Mannering, 2004; Yamamoto and Shankar, 2004; Khorashadi et al., 2005; Lee and Abdel-Aty, 2005; Islam and Mannering, 2006; Eluru et al., 2008; Savolainen and Mannering, 2007; Malyshkina and Mannering, 2010; Quddus et al., 2010).

However, the data intensive nature (requiring detailed and comprehensive data collection) of these traditional crash-severity models (their use of detailed crash-specific data) has been a considerable barrier to forecasting injury severities. To address this problem, Milton et al. (2008) applied a mixed logit model in their paper that used the injury outcome of the crash but not other detailed crash-specific data (such as driver characteristics, safety belt use, alcohol involvement). Their model looks at the proportion of crashes of each severity level on a specific roadway segment over a specified time period. With these proportions known, a standard crash-frequency model (one that models the total number of crashes regardless of severity) can be used in combination with the proportions model to determine the number of crashes by severity level without the need for detailed crash-specific data. The use of the mixed logit model in this application has advantages in that it can account for the unobserved heterogeneity that is likely to be present in using the limited data (roadway geometrics, pavement condition and general weather and traffic characteristics) as opposed to detailed crash-specific data.

However, the general question of how much information and accuracy is lost when not using detailed crash data to model injury severity remains. To provide some insight in this regard, this study provides a comparison of both fixed and random parameters (mixed) logit models using the traditional crash-injury severity approach (based on a crash having occurred and using detailed crash-specific data) and a crash-injury approach based only on non-crash specific data (other than observed injury severities) using crash-injury severity proportions as done in Milton et al. (2008).

2. Methodology

The application of the random parameters (mixed) logit model is undertaken by considering individual crash-injury severities (based on detailed crash-specific data) and by considering the proportions of crash-injury severities (proportions by injury severities) on specific roadway segments using only non-crash specific data (other than the crash's injury outcome). The proportions of fatality, injury and no-injury (property damage) on roadway segments (and the injury outcomes of specific crashes when individual crash severities are considered) are determined by the injury level of the most severely injured person in observed crashes. We follow Milton et al. (2008) and Washington et al. (2011), and start with

$$T_{in} = \boldsymbol{\beta}_i \mathbf{X}_{in} + \varepsilon_{in} \quad (1)$$

where T_{in} is a severity function determining the crash-injury severity category i on crash n for individual crash-data, and the proportion of crash-injury severity category i on roadway segment n for proportions crash-data; $\mathbf{X}_{in}$ is a vector of explanatory variables; $\boldsymbol{\beta}_i$ is a vector of estimable parameters for outcome i which may vary across observations, and ε_{in} the error term which is assumed to be generalized extreme value distributed (McFadden, 1981). To arrive at the mixed logit model, random parameters are introduced with $f(\boldsymbol{\beta}_i|\boldsymbol{\varphi})$, where $\boldsymbol{\varphi}$ is a vector of parameters of the density function (mean and variance). The resulting outcome probabilities are (see McFadden and Train, 2000; Train, 2003)

$$P_n(i|\boldsymbol{\varphi}) = \int \frac{e^{\boldsymbol{\beta}_i \mathbf{X}_{in}}}{\sum_{\forall I} e^{\boldsymbol{\beta}_i \mathbf{X}_{in}}} f(\boldsymbol{\beta}_i|\boldsymbol{\varphi}) d\boldsymbol{\beta}_i \quad (2)$$

where, $P_n(i|\boldsymbol{\varphi})$ is the outcome probability conditional on $f(\boldsymbol{\beta}_i|\boldsymbol{\varphi})$.[2] For model estimation, $\boldsymbol{\beta}_i$ can account for crash-specific variations of the effect of $\mathbf{X}$ on crash severity and crash injury-severity proportion probabilities, with the density function $f(\boldsymbol{\beta}_i|\boldsymbol{\varphi})$ used to determine $\boldsymbol{\beta}_i$. Mixed logit probabilities are then a weighted average for different values of $\boldsymbol{\beta}_i$ across crashes and roadway segments where some elements of the vector $\boldsymbol{\beta}_i$ may be fixed and some may be randomly distributed.

Estimation of the random parameters multinomial logit model shown in Eq. (2) can be undertaken using simulated maximum likelihood approaches, in which logit probabilities are approximated by drawing values of $\boldsymbol{\beta}_i$ from $f(\boldsymbol{\beta}_i|\boldsymbol{\varphi})$ for given values of $\boldsymbol{\varphi}$. Past research suggests that one of the best ways to draw values of $\boldsymbol{\beta}_i$ from $f(\boldsymbol{\beta}_i|\boldsymbol{\varphi})$ is to use a Halton sequence approach (for more on this technique, see: Halton, 1960; Bhat, 2003; Train, 2003). Research by Bhat (2003), Anastasopoulos and Mannering (2009), and others have shown that 200 Halton draws is usually sufficient for accurate parameter estimation (this number of Halton draws will be used in forthcoming model estimations). In this paper, for the functional form of the parameter density functions, consideration is given to normal, lognormal, triangular, uniform and Weibull distributions. With the functional forms of the parameter density functions specified, values of $\boldsymbol{\beta_i}$ are drawn from $f(\boldsymbol{\beta}_i|\boldsymbol{\varphi})$, logit probabilities are computed, and the simulated likelihood function is maximized.

3. Data

Crash data from rural interstate highways in Indiana were collected for a 5 year period (1995–1999 inclusive) from the Indiana Department of Transportation, Indiana State Patrol database, and Purdue's Center for Road Safety, to investigate the effect of pavement characteristics, highway geometries, traffic characteristics, driver socioeconomics and collision characteristics on crash injury severities and their proportions (these data were an expanded version of the crash-injury severity data previously used by Tarko et al., 2008). The severity outcomes considered are no-injury (property damage only), injury, and fatality.

The initial data sample consists of 5795 police-reported crashes that occurred on 231 freeway segments in Indiana with varying lengths (the mean segment length is approximately 1.15 miles). Of these 5795 police-reported crashes, 4658 resulted in no-injury, 1084 in injury, and 53 resulted in fatality. The 231 freeway segments were homogeneous (defined by roadway geometrics and pavement type). The roadway segment-defining information includes roadway geometrics, pavement characteristics, number of lanes, and speed limit. The data also include information on traffic and truck travel, and pavement condition.

To arrive at proportions injury data, over the 5-year data period, individual crash-data reports on the roadway segments are categorized based on the most severely injured person in the crash. Using this along with the total number of reported crashes per roadway segment, proportions of crash-injury severities are deter-

[2] Note that while the most common application of the multinomial logit model and its extensions (mixed logit, etc.) deal with discrete dependent variables, the multinomial logit framework is ideally suited to modeling fractional (proportions) dependent variables. In fact, the use of fractional dependent variables in the multinomial logit framework is standard code in many commercial software packages. See Papke and Wooldridge (1996) for a discussion of issues surrounding the modeling of fractional dependent variables.

Table 1
Descriptive statistics of explanatory variables.

	Mean	Std. dev.	Min	Max
Pavement characteristics				
Rut depth (in inches)	0.133	0.096	0.01	0.85
Friction number (on 0–100 scale)	38.112	9.241	17.9	67.4
International roughness index (in inches/mile)	82.248	32.440	27	264
Pavement condition rating (on 0–100 scale)	94.722	6.255	58	100
Geometric characteristics				
Road segment length (in miles)	1.154	1.708	0.057	11.53
Indicator variable: road segment length greater than 4 miles	0.310	0.463	0	1
Indicator variable: horizontal curve presence	0.791	0.407	0	1
Indicator variable: vertical curve presence	0.462	0.499	0	1
Number of vertical curves per lane-mile	0.229	0.520	0	5.618
Average vertical curve length per mile (in miles)	0.163	0.356	0	0.426
Number of ramps in the viewing direction	0.421	0.999	0	8
Interior shoulder width (in feet)	4.402	1.491	2.9	24.1
Indicator variable: interior shoulder width less than 4 feet	0.319	0.466	0	1
Outside shoulder width (in feet)	11.284	1.448	7.2	21
Indicator variable: outside shoulder width greater than 12 feet	0.224	0.417	0	1
Median width (in feet)	68.132	33.482	26	194.7
Indicator variable: median width greater than 65 feet	0.169	0.375	0	1
Indicator variable: median width less than 45 feet	0.237	0.425	0	1
Traffic characteristics				
Indicator variable: posted speed limit 55 miles/h	0.287	0.453	0	1
Average daily percent of trucks	0.340	0.108	0.088	0.449
Average daily percent of combination trucks	0.193	0.110	0.034	0.439
Driver and collision characteristics				
Indicator variable: driver age greater than 45 years	0.276	0.447	0	1
Indicator variable: driver had been drinking	0.034	0.181	0	1
Indicator variable: driver used no restraints	0.047	0.211	0	1
Indicator variable: driver illness/fatigued	0.018	0.132	0	1
Indicator variable: driver fall asleep	0.030	0.171	0	1
Indicator variable: driver not ejected	0.157	0.364	0	1
Indicator variable: type of the collision: angle	0.102	0.303	0	1
Indicator variable: type of the collision: head on	0.241	0.428	0	1
Indicator variable: light condition when the collision occurred: daylight	0.414	0.493	0	1
Indicator variable: light condition when the collision occurred: darkness – road lighted	0.026	0.160	0	1

mined over the 5-year period. These are used for the proportions crash-data model estimation.

Table 1 provides additional information on the mean, standard deviation, minimum and maximum of the explanatory variables.

4. Model estimation results

As shown in Table 2,[3] the individual crash-data mixed logit model had 24 variables that produced statistically significant parameters (two constants, rut depth (measured in inches), friction number (measured on 0–100 scale, with friction considered to be good if its value is 40 or above), road-segment length, road segment length greater than 4 miles, horizontal curve presence, vertical curve presence, number of vertical curves per mile, interior shoulder (in the direction of travel, the left-most, median lane) width, interior shoulder width less than 4 feet, median width, median width greater than 65 feet, posted speed limit 55 miles/h, average daily percent of combination trucks, driver age greater than 45 years, driver had been drinking, driver used no restraints, driver illness/fatigued, driver fell asleep, driver not ejected, angle collision, head-on collision, daylight at time of crash, darkness with road lighted at time of crash); 2 of these 24 variables (road segment length greater than 4 miles and median width greater than 65 feet) produced statistically significant random parameters (both were normally distributed). The proportions crash-data mixed logit model had 14 variables that produced statistically significant parameters (two constants, rut depth, friction number, international roughness index (measured in inches/mile, with lower values indicating smoother pavements), pavement condition rating (measured on 0–100 scale, with high PCR values indicating less deteriorated pavement), road segment length, outside shoulder (in the direction of travel, the shoulder next to the right-most lane) width greater than 12 feet, median width less than 45 feet – defined separately for no-injury and injury outcomes, number of ramps in the viewing direction, posted speed limit 55 miles/h, and percentage of trucks – defined separately for injury and fatality); 2 of these 14 variables (constant for no-injury and rut depth) produced statistically significant random parameters (both were normally distributed).[4]

Table 3 shows that both the individual and the proportions crash-data models have good overall statistical fit with McFadden pseudo-ρ^2 values in the 0.5–0.7 range. It is also noteworthy that the random parameter models (individual parameter estimation results shown in Table 2) provide a statistically superior fit relative to the traditional fixed-parameter models as indicated by the likelihood-ratio test results provided in Table 3 (although the proportions crash-data model has a much higher test statistic, as would be expected, due to there being less information than in the individual crash-data models, making random parameters more important).[5]

[3] To conserve space, the fixed-parameter model results are not presented.

[4] Model specifications were developed using t-statistics of individual parameters to ensure the significance of parameter estimates, and the improvement of the overall statistical fit of the model was assessed with likelihood ratio tests to make sure that each additional variable improves the overall statistical fit of the model. A 0.90 level of significance (p-value < 0.1) was generally used as a threshold to determine if parameter estimates for individual variables were significantly different from zero.

[5] Recent papers by Ye and Lord have explored the effect of sample size on the statistical fit of fixed and random parameter models (Ye and Lord, 2010a) and the effect of possible underreporting of minor crashes (Ye and Lord, 2010b). Future research could further investigate how underreporting of such minor crashes can be ultimately addressed with respect to model misspecifications, in the framework of the aggregate/disaggregate and fixed/random parameters analysis presented herein.

Table 2
Random parameters (mixed) logit model estimation results (F: fatality function; I: injury function; NI: no-injury function).

	Individual crash data mixed logit model		Proportions crash data mixed logit models	
	Parameter estimate	*t*-Ratio	Parameter estimate	*t*-Ratio
Constant [F]	−7.137	−3.788		
Constant [I]	4.259	1.682	1.839	2.364
Constant [NI]			6.084	3.516
(Std. dev. of parameter distribution – normally distributed) [NI]			(2.519)	(2.477)
Pavement characteristics				
Rut depth (in inches) [F and I]	−4.693	−2.288		
Rut depth (in inches) [I]			−3.702	−1.842
(Std. dev. of parameter distribution – normally distributed) [I]			(14.060)	(1.785)
Friction number (on 0–100 scale) [F and I]	−0.059	−2.443		
Friction number (on 0–100 scale) [F]			−0.058	−3.038
International roughness index (in inches/mile) [I]			0.026	1.830
Pavement condition rating (on 0–100 scale) [NI]			−0.062	−1.758
Geometric characteristics				
Road segment length (in miles) [I]			0.155	2.642
Road segment length (in miles) [NI]	−0.208	−2.087		
Indicator variable: road segment length greater than 4 miles [I]	−4.339	−2.898		
(Std. dev. of parameter distribution – normally distributed) [I]	(3.543)	(2.354)		
Indicator variable: horizontal curve presence [F and I]	−0.842	−2.251		
Indicator variable: vertical curve presence [F and I]	1.195	2.587		
Number of vertical curves per lane-mile [F and NI]	−0.940	−1.941		
Average vertical curve length per mile (in miles) [I]	−4.627	−3.005		
Interior shoulder width (in feet) [I]	−0.417	−2.562		
Indicator variable: interior shoulder width less than 4 feet [F and NI]	1.816	3.404		
Indicator variable: outside shoulder width greater than 12 feet [F]			1.024	2.173
Median width (in feet) [NI]	−0.051	−2.139		
Indicator variable: median width greater than 65 feet [I]	−3.162	−2.229		
(Std. dev. of parameter distribution – normally distributed) [I]	(2.668)	(1.792)		
Indicator variable: median width less than 45 feet [I]			−1.217	−3.593
Indicator variable: median width less than 45 feet [NI]			−0.892	−2.597
Number of ramps in the viewing direction [I]			−0.186	−2.739
Traffic characteristics				
Indicator variable: posted speed limit 55 miles/h [F and I]	1.610	3.380		
Indicator variable: posted speed limit 55 miles/h [F]			1.111	3.041
Percentage of trucks [F]			−1.774	−2.118
Percentage of trucks [I]			−2.075	−1.977
Average daily percent of combination trucks [NI]	−5.058	−2.691		
Driver and collision characteristics				
Indicator variable: driver age greater than 45 years [F]	2.445	2.127		
Indicator variable: driver had been drinking [I]	2.134	2.794		
Indicator variable: driver used no restraints [I]	2.743	4.312		
Indicator variable: driver illness/fatigued [I]	2.891	2.832		
Indicator variable: driver fall asleep [F and I]	1.713	2.606		
Indicator variable: driver not ejected [F and I]	−2.132	−2.587		
Indicator variable: type of the collision: angle [F and NI]	3.474	3.337		
Indicator variable: type of the collision: head on [F and NI]	1.368	2.938		
Indicator variable: light condition when the collision occurred: daylight [I]	0.782	2.413		
Indicator variable: light condition when the collision occurred: darkness – road lighted [I]	1.698	2.227		
Number of observations	5795		231	
Log-likelihood at zero, $LL(\mathbf{0})$	−744.14		−5683.11	
Log-likelihood at convergence, $LL(\boldsymbol{\beta})$	−250.33		−2718.71	

Table 3
Goodness-of-fit measures for the random and fixed parameter logit models.

	Individual crash data logit models		Proportions crash data logit models	
	Random parameters	Fixed parameters	Random parameters	Fixed parameters
Number of parameters	26	24	16	14
Log-likelihood at zero, $LL(\mathbf{0})$	−774.14	−774.14	−5683.11	−5683.11
Log-likelihood at convergence, $LL(\boldsymbol{\beta})$	−250.33	−266.13	−2718.71	−2788.40
McFadden pseudo-ρ^2	0.677	0.656	0.522	0.509
McFadden adjusted pseudo-ρ^2	0.643	0.625	0.519	0.507

Likelihood-ratio test	Random versus fixed parameters	Random versus fixed parameters
$X^2 = -2[LL(\boldsymbol{\beta}_{random}) - LL(\boldsymbol{\beta}_{fixed})]$	15.81	69.69
Degrees of freedom	2	2
Critical X^2 (0.999 level of confidence)	13.82	13.82
Number of observations	5795	231

Table 4
Observed versus model-predicted probabilities.

Model type	Observed versus model-predicted probabilities					
	Fatal		Injury		No-injury	
	Observed	Predicted	Observed	Predicted	Observed	Predicted
Fixed parameters individual crash data	0.0071	0.0094	0.1716	0.1787	0.8213	0.8119
Random parameters individual crash data		0.0071		0.1688		0.8241
Fixed parameters proportions crash data	0.0063	0.0120	0.1418	0.2219	0.8519	0.7660
Random parameters proportions crash data		0.0089		0.1850		0.8062

Table 5
Observed versus model-predicted injury-severity proportions of the 231 road segments predicted with the individual crash data and the proportions crash data.

Model type	Observed versus model-predicted injury-severity proportions of the 231 road segments					
	Fatal		Injury		No-injury	
	Observed	Predicted	Observed	Predicted	Observed	Predicted
Fixed parameters individual crash data	0.0063	0.0080	0.1418	0.1762	0.8519	0.8157
Random parameters individual crash data		0.0063		0.1555		0.8382
Fixed parameters proportions crash data		0.0120		0.2219		0.7660
Random parameters proportions crash data		0.0089		0.1850		0.8062

To further evaluate the statistical fit of the models, observed and predicted probabilities were compared. The results shown in Table 4 support the findings of Table 3 in that the random parameter and the individual crash-data models produced better observed versus model-predicted results (notice that, due to missing observations, the observed injury severity probabilities differ somewhat between individual crash-data and proportions data).

Table 5 provides a comparison of the observed and predicted injury-severity proportions for the 231 road segments, estimated with both the individual and the proportions crash -data models. For the estimation of the proportions with the individual crash-data models, the probabilities of all crashes on each of the 231 road segments were computed, and an average predicted probability was utilized as injury severity proportion for each segment. The results show that the individual crash-data models provide better predictors of the proportions compared to the proportions crash-data models, with the random parameters models outperforming their fixed parameters counterparts. This is further illustrated in Fig. 1, which presents the observed proportions, along with the range (the 10th and 90th percentile of the proportions) and mean of the predicted proportions by model type.

Although Table 5 shows that models using individual crash-data are better at predicting the mean of injury-severity proportions, Fig. 1 shows that the range of individual segment predictions tends to be tighter for proportions crash-data (see the injury and no-injury findings in Fig. 1) suggesting a smaller range of prediction error.

To provide additional insight, the measured error (the difference between observed and model-predicted probabilities) for fatalities,

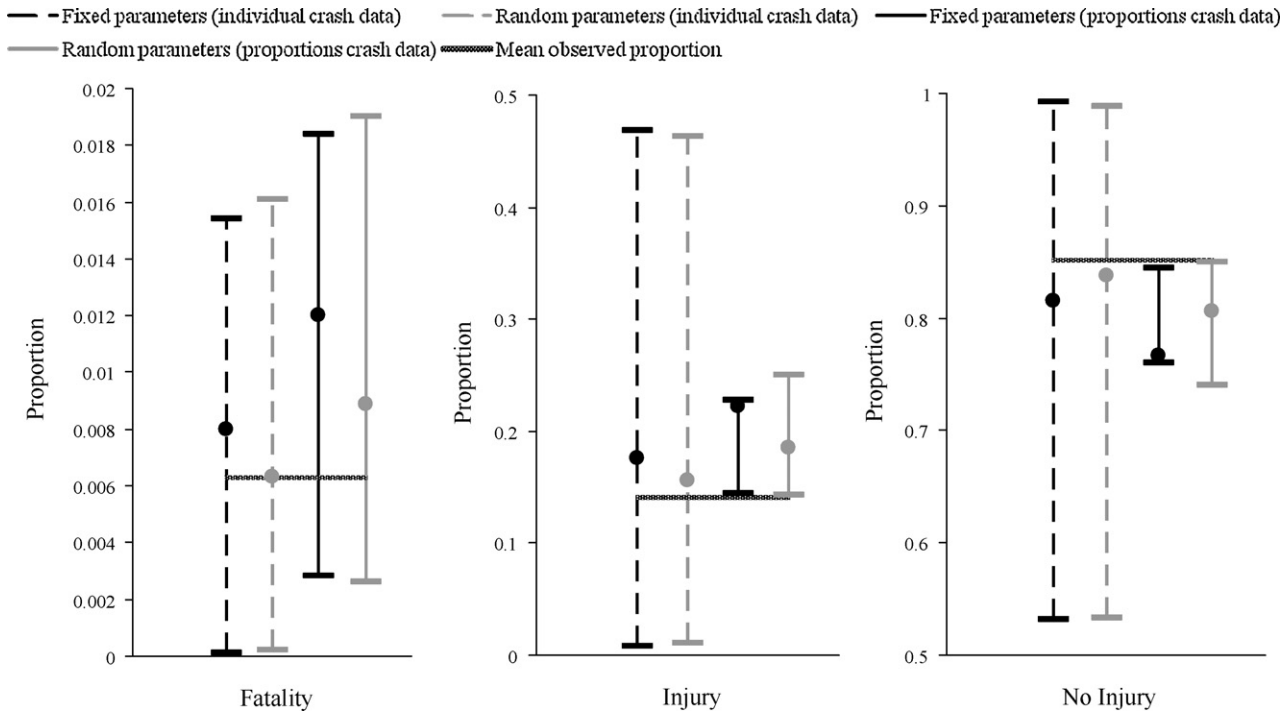

Fig. 1. Range (10th and 90th percentile) and means (dots) of the model-predicted proportions of the 231 road segments versus the mean observed proportions.

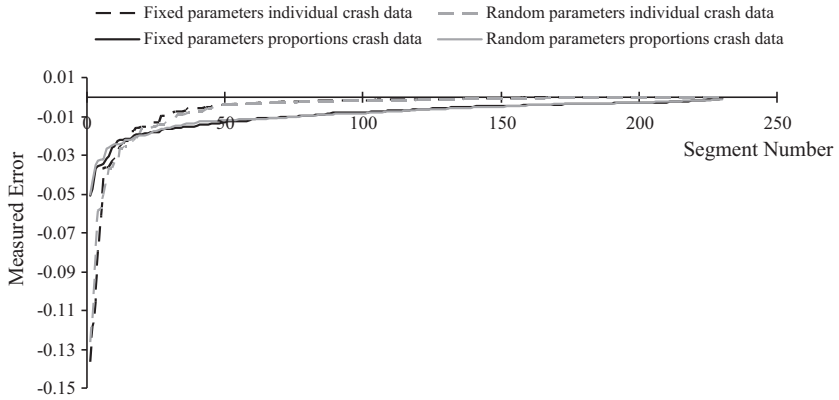

Fig. 2. Measured error (observed minus model-predicted proportions) of fatalities for 226 road segments (ordered from most negative to most positive). Note that 5 of the 231 roadway segments had high positive measurement error and were not included in this figure to allow more scale detail.

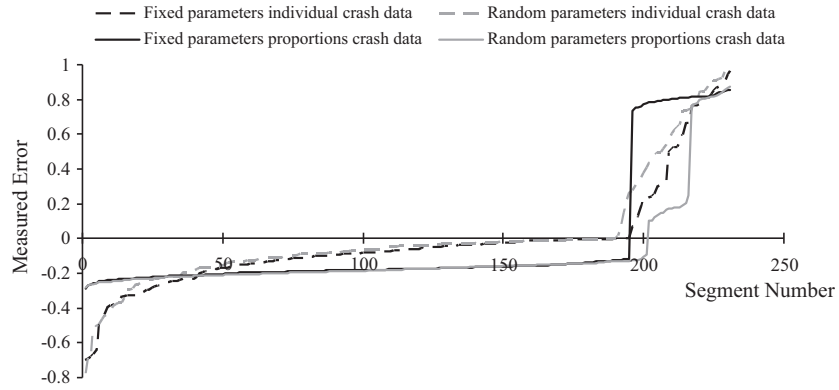

Fig. 3. Measured error (observed minus model-predicted proportions) of injuries for the 231 road segments (ordered from most negative to most positive).

injuries, and no-injuries for the 231 road segments is plotted in Figs. 2–4. These figures again show the superiority of the individual crash-data models, but also confirm Fig. 1 findings that proportions models seem less susceptible to extreme predictions (as indicated by their better performance in the tails of the curves shown in Figs. 2–4).

Another comparative assessment of the various models is how they might rank dangerous road segments. To do this, the predicted fatality proportions were determined on each segment for each of the four modeling approaches and the 12 segments with the highest predicted proportions of fatalities (roughly 5% of the 231 highway segments) were identified. Using the random-parameters model on individual-crash data as the base (because this model contains the most information), we find that the random parameters model based on proportions data identifies 11 of the 12 segments determined to have the highest fatality proportions by the random-parameters model on individual-crash data – suggesting that little practical information is lost using just non-crash specific data (both the fixed parameters individual crash model and the proportions crash-data model had 9 of the 12 segments in common with the random parameters individual crash-data model).

From all of this we can infer that the individual crash-data models fit the data better than the proportions crash-data models – an expected result given the more limited data used in the

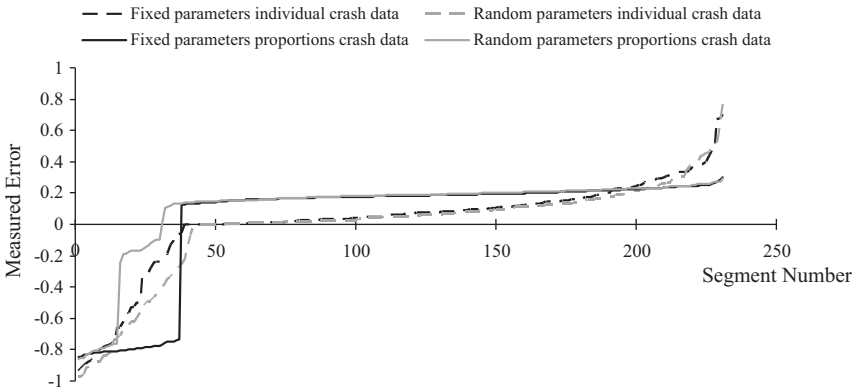

Fig. 4. Measured error (observed minus model-predicted proportions) of no-injuries for the 231 road segments (ordered from most negative to most positive).

proportions models. However, the loss in accuracy is rather modest as indicated by the dangerous-segment ranking result and the results in Tables 4 and 5. And individual crash-data models seem to be somewhat more susceptible to extreme predictions (see Figs. 1–4) for injury and no-injury severity proportions. In choosing one approach over the other, careful consideration must be given to data requirements and the intended use of the model. If the intended use of the model is for forecasting injury severities, the proportions model provides reasonable accuracy and has explanatory variables that are more aggregate and easier to project. If the intended use is to uncover fundamental relationships of injury severities, individual crash-data models provide a much richer set of information due to the fact that information is drawn from detailed crash data.

5. Summary and conclusion

In efforts that seek to reduce crash frequencies and severities, it is common practice to identify roadway segments that have a high number of severe-crash occurrences. This paper provides a comparison of fixed and random parameters logit models using two types of injury-severity data: one using detailed crash-specific, individual-crash data and the other using proportions data that considers the proportion of crashes by severity level for specified roadway segments.

The model comparison showed the statistical superiority of the random-parameter logit model compared to the fixed-parameter logit model. The findings also show that the models based on individual crash-data provide better overall fit relative to the models based on the proportions of crashes by severity type. However, the choice between individual crash-data and proportions crash-data models is not necessarily obvious because they both have their merits and limitations. Models based on individual crash-data can provide fundamental insights into injury-severity determination using detailed data from crash reports. However, these models can be cumbersome to use for estimating changes in injury severities resulting from safety countermeasures due to the many explanatory variables that need to be determined. Models based on the proportions of crash injuries on specific roadway segments use aggregate data on roadway, traffic and weather conditions making estimating changes in severity levels much less cumbersome. It is noteworthy that our model assessments suggest that using proportions data may be practically very close to using individual-crash data because both approaches were found to identify nearly the same roadway segments ranked in the top 5% by highest fatality proportions. However, it is important to keep in mind that the aggregate nature of the data used in the proportions-data models does limit the fundamental insights that can be gathered from the modeling process.

Acknowledgements

The authors would like to thank Professor Andrew Tarko, from the Center for Road Safety, for providing the driver– and collision-specific data and for his helpful suggestions and comments, Natalie Villwock for her help with data collection, and Bill Flora for providing the pavement characteristics datasets.

References

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

Aguero-Valverde, J., Jovanis, P., 2009. Bayesian multivariate Poisson log-normal models for crash severity modeling and site ranking. In: Paper presented at the 88th Annual Meeting of the Transportation Research Board , Washington, DC.

Anastasopoulos, P.Ch., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153–159.

Bijleveld, F.D., 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis and Prevention 37 (4), 591–600.

Bhat, C., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transportation Research Part B 37 (1), 837–855.

Bonneson, J.A., Pratt, M.P., 2008. Procedure for developing accident modification factors from cross-sectional data. Transportation Research Record 2083, 40–48.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on accident frequencies and severities. Accident Analysis and Prevention 33 (1), 99–109.

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.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40 (4), 1257–1266.

Duncan, C., Khattak, A., Council, F., 1998. Applying the ordered probit model to injury severity in truck–passenger car rear-end collisions. Transportation Research Record 1635, 63–71.

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41 (4), 820–828.

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

Geedipally, S.R., Lord, D., 2010. Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson–Gamma models. Accident Analysis and Prevention 40 (3), 1123–1134.

Halton, J., 1960. On the efficiency of evaluating certain quasi-random sequences of points in evaluating multi-dimensional integrals'. Numerische Mathematik 2 (1), 84–90.

Islam, S., Mannering, F., 2006. Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. Journal of Safety Research 37 (3), 267–276.

Khattak, A., 2001. Injury severity in multi-vehicle rear-end crashes. Transportation Research Record 1746, 59–68.

Khattak, A., Pawlovich, D., Souleyrette, R., Hallmark, S., 2002. Factors related to more severe older driver traffic crash injuries. Journal of Transportation Engineering 128 (3), 243–249.

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

Kockelman, K., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34 (4), 313–321.

Kweon, Y.-J., Kockelman, K., 2003. Overall injury risk to different drivers: combining exposure, frequency, and severity models. Accident Analysis and Prevention 35 (3), 414–450.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34 (2), 149–161.

Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accident Analysis and Prevention 37 (4), 775–786.

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

Ma, J., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models of injury count by severity. Transportation Research Record 1950, 24–34.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964–975.

Malyshkina, N., Mannering, F., 2010. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accident Analysis and Prevention 42 (1), 131–139.

McFadden, D., 1981. Econometric models of probabilistic choice. In: Manski, D., McFadden (Eds.), A Structural Analysis of Discrete Data with Econometric Applications. The MIT Press, Cambridge, MA.

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15 (5), 447–470.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus Empirical Bayes. Transportation Research Record 1840, 31–40.

Miaou, S.-P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis and Prevention 37 (4), 699–720.

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

Papke, L., Wooldridge, J., 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. Journal of Applied Econometrics 11 (6), 619–632.

Park, E.-S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record 2019, 1–6.

Quddus, M.A., Wang, C., Ison, S., 2010. Road traffic congestion and crash severity: an econometric analysis using ordered response models. Journal of Transportation Engineering 136 (5), 424–435.

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.

Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accident Analysis and Prevention 28 (3), 391–401.

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97 (1), 246–273.

Tarko, A.P., Villwock, N.M., Blond, N., 2008. Effect of median design on rural freeway safety: flush medians with concrete barriers and depressed medians. Transportation Research Record: Journal of the Transportation Research Board 2060, 29–37.

Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, UK.

Ulfarsson, G., Mannering, F., 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accident Analysis and Prevention 36 (2), 135–147.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2011. Statistical and Econometric Methods for Transportation Data Analysis, second ed. Chapman & Hall/CRC.

Yamamoto, T., Shankar, V., 2004. Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. Accident Analysis and Prevention 36 (5), 869–876.

Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Science 47 (3), 443–452.

Ye, F., Lord, D., 2010a. Comparing Three Commonly used Crash Severity Models on Sample Size Requirements: Multinomial Logit, Ordered Probit and Mixed Logit Models. Working Paper, Texas A&M University, College Station, TX.

Ye, F., Lord, D., 2010b. Investigating the Effects of Underreporting of Crash Data on Three Commonly used Traffic Crash Severity Models: Multinomial Logit, Ordered Probit and Mixed Logit Models. Working Paper, Texas A&M University, College Station, TX.