

Crash Injury Severity Analysis Using Bayesian Ordered Probit Models

Yuanchang Xie¹; Yunlong Zhang²; and Faming Liang³

Abstract: Understanding the underlying relationship between crash injury severity and factors such as driver's characteristics, vehicle type, and roadway conditions is very important for improving traffic safety. Most previous studies on this topic used traditional statistical models such as ordered probit (OP), multinomial logit, and nested logit models. This research introduces the Bayesian inference and investigates the application of a Bayesian ordered probit (BOP) model in driver's injury severity analysis. The OP and BOP models are compared based on datasets with different sample sizes from the 2003 National Automotive Sampling System General Estimates System (NASSGES). The comparison results show that these two types of models produce similar results for large sample data. When the sample data size is small, with proper prior setting, the BOP model can produce more reasonable parameter estimations and better prediction performance than the OP model. This research also shows that the BOP model provides a flexible framework that can combine information contained in the data with the prior knowledge of the parameters to improve model performance.

DOI: 10.1061/(ASCE)0733-947X(2009)135:1(18)

CE Database subject headings: Bayesian analysis; Monte Carlo method; Traffic accidents; Injuries.

Introduction

The relationship between the injury severity of traffic crashes and factors such as driver and passenger characteristics, vehicle type, and traffic and geometric conditions has attracted much attention. Better understanding of this relationship is necessary and very important for improving vehicle and roadway designs such that severe injuries can be reduced. Numerous studies have applied statistical models for crash injury severity study. Among them, the ordered probit (OP), ordered logit and their variations are the most often used models. O'Donnell and Connor (1996) applied an OP model to predict the injury severity of motor vehicle occupants. Duncan et al. (1998) used an OP model to model the injury severity of passenger car occupants in rear-end crashes between trucks and passenger cars. Klop and Khattak (1999) examined factors affecting the injury severity to bicyclists, also using the OP model. In recent years, the OP model has been widely accepted and used for crash injury severity studies by many transportation safety researchers, including Kockelman and Kweon (2002), Abdel-Aty (2003), Khattak et al. (2002), Zajac and Ivan (2003), Khattak and Targa (2004), Abdel-Aty and Abdelwahab (2004), Dissanayake and Ratnayake (2005), Lee and Abdel-Aty

(2005), Riffat and Chor (2005), and Siddiqui et al. (2006). In addition, a variation of the OP model called the bivariate ordered probit model was also used by Yamamoto and Shankar to analyze the injury severity of both the driver and the most severely injured passenger (Yamamoto and Shankar 2004).

Besides the OP model, several other models have also been used. These models include ordered logit model (O'Donnell and Connor 1996), multinomial logit model (Khorashadi et al. 2005; Savolainen and Mannering 2007), nested logit model (Shankar et al. 1996; Chang and Mannering 1999; Savolainen and Mannering 2007), mixed logit model (Srinivasan 2002; Milton et al. 2008), heteroscedastic ordered logit model (Wang and Kockelman 2005), and logistic regression (Al-Ghamdi 2002). There is no consensus on which model is the best. Several researchers argued that categorical models such as the multinomial logit model may be better than the ordered logit and the OP models in that the ordered models restrict the effect of variables across outcomes (Khorashadi et al. 2005). Abdel-Aty (2003) applied the OP model, multinomial logit model, and the nested logit model to injury severity analysis for roadway sections, signalized intersections, and toll plazas, and concluded that the multinomial model produced poorer results and lower goodness-of-fit measure than the OP model. In addition, he examined the performance of six nested logit models and found that for all geometric classes, the best performing nested logit model almost did not improve the classification accuracy of the OP model. Although the goodness-of-fit measure was slightly improved, fewer variables were found to be significant and it was more difficult to implement the nested logit model compared with the OP model.

For all the studies discussed in the previous two paragraphs, the model fittings are mostly based on the maximum likelihood estimation (MLE) method regardless of the models used. MLE is a standard method that has been widely used for model fitting. However, the model fitting result from the MLE method depends completely on the quality of the data. If the data cannot adequately represent characteristics of the population, which is often the case when the size is small, the fitted model will most likely

¹Assistant Professor, Dept. of Civil and Mechanical Engineering Technology, South Carolina State Univ., P.O. Box 8144, 300 College ST NE, Orangeburg, SC 29117. E-mail: yxie@scsu.edu

²Assistant Professor, Zachry Dept. of Civil Engineering, Texas A&M Univ., 3136 TAMU, College Station, TX 77843 (corresponding author). E-mail: yzhang@civil.tamu.edu

³Associate Professor, Dept. of Statistics, Texas A&M Univ., 3143 TAMU, College Station, TX 77843. E-mail: fliang@stat.tamu.edu

Note. Discussion open until June 1, 2009. Separate discussions must be submitted for individual papers. The manuscript for this paper was submitted for review and possible publication on May 30, 2007; approved on July 2, 2008. This paper is part of the *Journal of Transportation Engineering*, Vol. 135, No. 1, January 1, 2009. ©ASCE, ISSN 0733-947X/2009/1-18-25/\$25.00.

be erroneous and misleading. Although for some traffic safety studies, it is possible to obtain large sample data (Chang and Mannering 1999; Kockelman and Kweon 2002; Milton et al. 2008), for many transportation safety studies with specific purposes, it is costly, or sometimes even impossible, to obtain large sample data (Lord and Bonneson 2005).

The objective of this paper is to introduce Bayesian inference into injury severity modeling of traffic crashes. Compared with the traditional models that use the MLE method for parameter estimation, the Bayesian inference provides a flexible framework that can incorporate analysts' prior knowledge of the data such that the model fitting is not completely dependent on the data. This advantage of using Bayesian inference will be further detailed later in the paper. Readers can also refer to O'Hagan and Luce (2003) for more benefits of using Bayesian inference. In this research, a Bayesian ordered probit (BOP) model is introduced and applied to analyze drivers' injury severity for data of both large and small sample sizes. The OP model is chosen as the baseline for comparison mainly due to its popularity.

Methodological Background

OP Model

The OP model, also known as the ordinal probit model, is commonly used for analyzing data sets that include categorical and ordered dependent variables. For example, the traffic crash injury severity is normally labeled categorically as "no injury," "no injury but complaint of pain," "nonincapacitating injury," "incapacitating injury," and "fatal" (Duncan et al. 1998). In the 2003 National Automotive Sampling System General Estimates System (NASSGES) used in this study, those five categories are also used for crash injury severity (NHTSA 2003). Since there is an underlying ordering of these labels, ordered models such as the OP model are often used to model the subject. Although the OP model has been described in many references, to make this paper self-contained, a brief introduction of the OP model is presented here. More information can be found in the following references: Mckelvey and Zavoina (1975), Albert and Chib (1993) and Johnson and Albert (1999).

Crash injury severity is believed to be related to a number of factors including drivers and occupants characteristics, vehicle type, and geometric conditions. Let y_i represent the injury severity that has C categories, and \mathbf{x}_i represent variables that may affect the injury severity. Next, a latent variable z_i is introduced as

$$z_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \quad (1)$$

where $\mathbf{x}_i = \{1, x_{i1}, \dots, x_{ij}, \dots, x_{im}\}^T$ = input value for the i th individual ($i = 1, \dots, n$ and $j = 1, \dots, m$); n = total number of observations; m = total number of variables; x_{ij} = value of the j th variable for the i th individual; $\boldsymbol{\beta} = \{\beta_0, \beta_1, \dots, \beta_j, \dots, \beta_m\}^T$ = column vector of coefficients; and ε_i = random error term following standard normal distribution.

The value of the dependent variable y_i is then determined as

$$y_i = \begin{cases} 1 & \text{if } \gamma_0 < z_i \leq \gamma_1 \\ k & \text{if } \gamma_{k-1} < z_i \leq \gamma_k \\ C & \text{if } \gamma_{C-1} < z_i \leq \gamma_C \end{cases} \quad (2)$$

where $\boldsymbol{\gamma} = \{\gamma_0, \dots, \gamma_k, \dots, \gamma_C\}$ = threshold values for all categories. γ_0 and γ_C are defined as $-\infty$ and $+\infty$, respectively. The remaining threshold values are subject to constraint $\gamma_1 \leq \dots \leq \gamma_k \leq \dots$

$\leq \gamma_{C-1}$. Given the value of \mathbf{x}_i , the probability that the injury severity of individual i belongs to each category is

$$P(y_i = 1) = \Phi(\gamma_1 - \mathbf{x}_i^T \boldsymbol{\beta})$$

$$P(y_i = k) = \Phi(\gamma_k - \mathbf{x}_i^T \boldsymbol{\beta}) - \Phi(\gamma_{k-1} - \mathbf{x}_i^T \boldsymbol{\beta})$$

$$P(y_i = C) = 1 - \Phi(\gamma_{C-1} - \mathbf{x}_i^T \boldsymbol{\beta}) \quad (3)$$

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

To make the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}' = \{\gamma_1, \dots, \gamma_{C-1}\}$ identifiable (Mckelvey and Zavoina 1975; Albert and Chib 1993), one restriction is imposed on the threshold values to make $\gamma_1 = 0$. The unknown parameters needing to be estimated then become $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}^* = \{\gamma_2, \dots, \gamma_{C-1}\}$. For the classical OP model, the values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}^*$ can be determined by the MLE method. It is easy to show that the likelihood function of the data is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}^* | \mathbf{y}) = \prod_{i=1}^n \prod_{k=1}^C [\Phi(\gamma_k - \mathbf{x}_i^T \boldsymbol{\beta}) - \Phi(\gamma_{k-1} - \mathbf{x}_i^T \boldsymbol{\beta})]^{I(y_i=k)} \quad (4)$$

where $I(y_i=k)$ = indicator function. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}^*$ can be determined by maximizing $L(\boldsymbol{\beta}, \boldsymbol{\gamma}^* | \mathbf{y})$. However, there are several limitations with this MLE method. For example, if the data cannot represent the population, the estimated model may be erroneous, since the parameter estimation results depend completely on the data. In addition, the maximization process is a nonlinear optimization problem, which is not guaranteed to converge to a global optimal solution (Mckelvey and Zavoina 1975).

Bayesian Inference and BOP Model

Bayesian Inference

For Bayesian inference, the parameters to be estimated are assumed to follow certain prior distributions. These prior distributions reflect analysts' prior knowledge about the data to be analyzed. If the current knowledge of the data cannot be summarized in the form of an informative prior, some noninformative priors will usually be used (Congdon 2003). Based on the data, the likelihood function is used to update the prior distributions and obtain the posterior distribution of parameters. Let $\boldsymbol{\theta}$ denote the parameters to be estimated, the posterior distribution of $\boldsymbol{\theta}$ is then calculated as

$$\pi(\boldsymbol{\theta} | \mathbf{y}) = \frac{f(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{m(\mathbf{y})} \propto f(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \quad (5)$$

where $\mathbf{y} = \{y_1, \dots, y_i, \dots, y_n\}$ = observed outcomes; $\pi(\boldsymbol{\theta})$ = prior distribution of $\boldsymbol{\theta}$; $f(\mathbf{y} | \boldsymbol{\theta})$ = sampling distribution; $m(\mathbf{y}) = \int_{\boldsymbol{\theta}} f(\mathbf{y} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ = marginal distribution of \mathbf{y} ; and $\pi(\boldsymbol{\theta} | \mathbf{y})$ = posterior distribution of $\boldsymbol{\theta}$.

From Eq. (5), one can see that the Bayesian inference provides a flexible framework such that the prior knowledge of the data can be incorporated into the parameter estimation process.

BOP Model

Based on the previous discussions on likelihood function and Bayesian inference, it is easy to extend the Bayesian inference into the OP model. To facilitate the implementation of the BOP model, a data augmentation technique (Albert and Chib 1993; Johnson and Albert 1999) is used such that the latent variables \mathbf{z} are treated as unknown parameters to be estimated, and the final joint posterior distribution for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}^*$, and \mathbf{z} is shown in Eq. (6)

Table 1. Description of Variables Used in This Study

Variable type	Variable name	Description
Dependent variable	INJSEV	Injury severity level: 0=no injury, 1=possible injury, 2=nonincapacitating injury, 3=incapacitating injury, and 4=fatal injury
Driver info.	AGE	Age of driver (year)
	GENDER	0=female, and 1=male
	ALCOHOL	0=alcohol not involved, and 1=alcohol involved
Vehicle info.	SUV	Driver driving a sport utility vehicle (0=no, and 1=yes)
	VAN	Driver driving a van (0=no, and 1=yes)
	VAGE	Vehicle age (2004—model year)
Crash info.	ROLLOVER	If there was rollover happened: (0=no, and 1=yes)
	FIRE	If there was fire involved: (0=no, and 1=yes)
	FRONT	The initial impact point is front, front left, or front right: (0=no, and 1=yes)
	RIGHT	The initial impact point is right: (0=no, and 1=yes)
	LEFT	The initial impact point is left: (0=no, and 1=yes)
	BACK	The initial impact point is back, back left, or back right: (0=no, and 1=yes)
	TOP	The initial impact point is top: (0=no, and 1=yes)
Road info.	JUNC	Crash related to junctions: (0=no, and 1=yes)
	ALIGN	Roadway alignment: (0=straight, and 1=curve)
	PROF	Roadway profile: (0=level, and 1=others)
	SURF	Roadway surface condition: (0=dry, and 1=others)
	LIGHT	Light condition at the time of the crash: [0=daylight (excluding dusk and dawn), and 1=others]
	WEATH	Weather condition when the crash happened: (0=no adverse weather conditions, 1=others)

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}^*, \mathbf{z} | \mathbf{y}) \propto \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}^*) \prod_{i=1}^n \left\{ \phi(z_i - \mathbf{x}_i^T \boldsymbol{\beta}) \times \prod_{k=1}^C [I(\gamma_{k-1} < z_i < \gamma_k)]^{I(y_i=k)} \right\} \quad (6)$$

where $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}^*)$ =prior distribution of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}^*$; $\phi(\cdot)$ =probability density function of the standard normal distribution; and $\mathbf{z}=\{z_1, \dots, z_n\}$ and $z_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, 1)$.

Markov chain Monte Carlo (MCMC) algorithms can be used to simulate approximately independent samples from this joint posterior distribution. For complete information about the MCMC algorithm and how to obtain independent samples from the posterior distribution, please refer to Johnson and Albert (1999). In this research, the MCMC algorithm proposed by Cowles (1996) is employed.

Comparison between OP Model and BOP Model

As can be seen from Eqs. (4) and (6), for the OP model, the estimation of parameters depends completely on the data, while for the BOP model additional prior distributions are included for estimating the parameters. The prior distributions provide extra flexibility that makes it possible to incorporate analysts' knowledge of the parameters into the estimation process. This is especially important for data with small sample sizes that may not adequately represent population characteristics.

Data Source and Test Design

Data Source

The focus of this research is to investigate the injury severity of drivers in motor vehicle crashes. Data were obtained from the 2003 NASSGES, and only crash records related to automobiles, SUVs, and vans were used. There were a total of 76,994 usable records in the database. The driver injury severity was classified into five categories: no injury, possible injury, nonincapacitated injury, incapacitated injury, and fatal injury. For the dataset used, these categories have 50,347, 10,673, 7,730, 7,577, and 667 observations, respectively. It can be seen that more than half of the drivers involved in crashes were not injured. There are 19 independent variables used to explain the injury severity of drivers. These variables are listed in Table 1.

Test Design

Based on the data, an OP model is first fitted. BOP models with noninformative and informative prior distributions are also fitted and compared with the OP model. For the OP model, the likelihood ratio index (LRI), sometimes also referred to as McFadden's pseudo R^2 , is provided to measure the goodness-of-fit of the model. The LRI has often been used to evaluate discrete choice models (Klop and Khattak 1999; Kockelman and Kweon 2002; Abdel-Aty 2003). In addition, the likelihood ratio value is also calculated.

To investigate the effect of small sample size on the model

Table 2. Estimation Result of OP Model

Parameter	Mean	Std.	<i>t</i> value	Approx. Pr> <i>t</i>
Intercept	-0.7778	0.0366	-21.23	<0.0001
AGE	0.0032	0.0003	12.18	<0.0001
GENDER	-0.2067	0.0090	-23.01	<0.0001
ALCOHOL	0.3311	0.0195	17.02	<0.0001
SUV	-0.1534	0.0125	-12.28	<0.0001
VAN	-0.1624	0.0159	-10.21	<0.0001
VAGE	0.0121	0.0009	14.25	<0.0001
ROLLOVER	1.1161	0.0225	49.55	<0.0001
FIRE	0.9621	0.0789	12.19	<0.0001
FRONT	0.2533	0.0337	7.51	<0.0001
RIGHT	0.2109	0.0351	6.01	<0.0001
LEFT	0.3719	0.0349	10.66	<0.0001
BACK	0.1885	0.0350	5.38	<0.0001
TOP	0.2244	0.2225	1.01	0.3133
JUNC	-0.0757	0.0093	-8.14	<0.0001
ALIGN	0.1267	0.0137	9.21	<0.0001
PROF	0.0031	0.0107	0.29	0.7740
SURF	-0.0739	0.0161	-4.59	<0.0001
LIGHT	0.1292	0.0100	12.94	<0.0001
WEATH	-0.0147	0.0187	-0.79	0.4319
γ_2	0.4424	0.0040	111.11	<0.0001
γ_3	0.8977	0.0059	151.44	<0.0001
γ_4	2.1280	0.0156	136.62	<0.0001
Likelihood ratio	—	5,400.3	—	—
LRI	—	0.0333	—	—
Number of observations	—	76,994	—	—
Log-likelihood at zero	—	-80,980	—	—
Log-likelihood at convergence	—	-78,280	—	—

fitting and prediction performances, several small datasets are randomly sampled from the original data. The small sample data are used to fit both the OP model and the BOP model, and the fitted models are then applied to the remaining data to test their prediction performance.

Analysis of Results

Application to Original Data

Results from OP Model

Based on the 76,994 crash records, the OP model was fitted using the SAS (SAS Institute Inc. 2004) software package. The fitted results are listed in Table 2. For the likelihood ratio test, the null hypothesis is that all the parameters of the independent variables are zeros. In our case, the likelihood ratio value is 5,400.3 (see Table 2). Since there are 19 variables, at the 0.05 significance level the critical χ^2 value is $\chi_{0.05}^2(19)=30.1$, which is much less than the calculated likelihood ratio value. This suggests that the null hypothesis should be rejected at the 0.05 significance level. Large LRI value means better model fit. The calculated LRI value is 0.0333, which is comparable with the values reported in similar studies (Klop and Khattak 1999; Kockelman and Kweon 2002).

Table 2 also shows the estimated parameters for each independent variable, the intercept, and the threshold values for each injury severity category. Since γ_0 , γ_1 , and γ_5 are set to be $-\infty$, 0,

and $+\infty$, respectively, only values of γ_2 , γ_3 , and γ_4 are estimated and shown in Table 2. The signs for the parameters of the independent variables reflect the variables' effects on driver's injury severity. A positive sign means an increase of the value of the variable will increase the probability of the most severe category of injury and decrease the probability of the least severe category of injury (Washington et al. 2003).

The estimated parameters for the independent variables are generally consistent with the results reported by Kockelman and Kweon (2002), and Yamamoto and Shankar (2004). The positive values for AGE and VAGE suggest that senior drivers or drivers in older vehicles tend to have slightly more severe injuries. The negative sign for GENDER means that the chance for male drivers to suffer the most severe category of injuries is less than female drivers under the same crash circumstances. The high parameter value for variable ALCOHOL shows that drunk driving is very dangerous and significantly increases the possibility of more severe injuries. The fitted results also indicate that compared to automobiles, SUVs and vans can better protect drivers given a crash has happened (note that we only considered three vehicle types: automobiles, SUVs, and vans. Thus two dummy variables, SUV and VAN, are enough). ROLLOVER and FIRE appear to be the two most critical factors that significantly aggravate driver's injury severity level. The effects of initial impact points on driver's injury severity are also evaluated and the results show that the initial impact points on the left side are the most dangerous. This conclusion is intuitive, since in general the closer the driver is to the initial impact point, the worse the injury severity would be to the driver. Other parameters show that curvy road alignments and inadequate light conditions may result in more severe injuries. It is interesting to find out that variables JUNC, SURF, and WEATH have negative signs, which means crashes related to junctions or interchanges, icy or wet surfaces, and adverse weather can actually lead to lower probability of suffering the most severe category of injuries. These findings may be viewed as counterintuitive, but Yamamoto and Shankar (2004) also reported similar results in their study. A possible explanation is that under such conditions, drivers tend to drive at lower speeds and be more cautious.

Results from BOP Model

To compare with the OP model, three BOP models with different prior distributions were also applied to the same 76,994 crash records. For the first BOP model, uniform prior distributions were chosen for all parameters to be estimated. In this case, almost no information was provided in the prior distributions, also termed noninformative prior distributions; for the second BOP model, all prior distributions were assumed to be normally distributed with zero mean and a variance value of 0.5; for the third BOP model, all prior distributions were assumed to be normally distributed with zero mean and a variance value of 0.1. Actually, the first BOP model can be considered as using normal prior distributions with zero mean and infinite variance. For all three BOP models, the prior distributions of the threshold values (γ^*) were set to be uniformly distributed. With the decrease in the variance value, the prior distributions became increasingly informative, which means the analysts were more confident about the prior distributions. The selections of the mean values were simplified and rather arbitrary for this application, and more complicated selections of the mean values were considered for the data with small sample size in the next section.

Using MCMC simulation, samples from the posterior distribution of each parameter can be obtained. From these samples, an

Table 3. Parameter Estimation Results of BOP Model

Parameter	First BOP model		Second BOP model		Third BOP model	
	Mean	Std.	Mean	Std.	Mean	Std.
Intercept	-0.7785	0.0363	-0.7754	0.0384	-0.7508	0.0390
AGE	0.0032	0.0003	0.0032	0.0003	0.0032	0.0003
GENDER	-0.2066	0.0084	-0.2070	0.0092	-0.2065	0.0099
ALCOHOL	0.3314	0.0205	0.3278	0.0188	0.3292	0.0180
SUV	-0.1535	0.0120	-0.1542	0.0129	-0.1543	0.0119
VAN	-0.1633	0.0166	-0.1623	0.0150	-0.1637	0.0157
VAGE	0.0121	0.0009	0.0122	0.0009	0.0121	0.0009
ROLLOVER	1.1142	0.0235	1.1140	0.0231	1.1056	0.0226
FIRE	0.9647	0.0784	0.9503	0.0806	0.9090	0.0771
FRONT	0.2544	0.0336	0.2519	0.0342	0.2299	0.0344
RIGHT	0.2105	0.0345	0.2098	0.0357	0.1882	0.0357
LEFT	0.3718	0.0349	0.3706	0.0342	0.3485	0.0364
BACK	0.1890	0.0339	0.1855	0.0359	0.1641	0.0356
TOP	0.3135	0.2159	0.2564	0.2149	0.1835	0.1698
JUNC	-0.0749	0.0092	-0.0756	0.0094	-0.0764	0.0095
ALIGN	0.1274	0.0144	0.1260	0.0135	0.1270	0.0152
PROF	0.0035	0.0109	0.0035	0.0097	0.0008	0.0102
SURF	-0.0757	0.0155	-0.0769	0.0165	-0.0753	0.0169
LIGHT	0.1303	0.0093	0.1301	0.0101	0.1288	0.0097
WEATH	-0.0129	0.0194	-0.0120	0.0186	-0.0134	0.0191
γ_2	0.4426	0.0039	0.4429	0.0041	0.4425	0.0038
γ_3	0.8984	0.0055	0.8982	0.0058	0.8977	0.0055
γ_4	2.1288	0.0157	2.1288	0.0156	2.1272	0.0151

approximate density function can be drawn for each parameter and the posterior mean and standard deviation values can also be estimated. Table 3 shows the fitted results of the three BOP models. One can see that the estimated means from the OP model and the first BOP model are almost identical. One possible reason for this similarity is the noninformative uniform prior distributions used by the first BOP model. Since the noninformative prior distributions did not really provide any useful information, the Bayesian method could only use the information contained in the data. Another reason for this similarity might be due to data characteristics. It is possible that the data do not have a unique distribution and therefore the Bayesian method with noninformative distribution provides the same parameter estimations. Another finding is that with the decrease of the prior variances, most of the estimated parameters become closer to zeros and their standard deviations also become smaller. This is expected since smaller prior variances mean stronger confidence in the prior distributions and consequently lead to narrower credible intervals. Overall, the difference between the OP and BOP models in this case is not significant. This is because the sample size is very large. When sample size is large enough, Bayesian and the MLE methods will generally produce similar results.

Application to Small Sample Data

Sample size can significantly affect the accuracy of parameter estimation of regression models. This problem is commonly known as the consistency property of parameter estimators in the field of econometrics. With increased sample sizes, a good estimator should have a higher probability of being closer to the parameter it estimates while small sample sizes usually have the opposite effect on parameter estimation (Greene 2000; Washington et al. 2003). One such example related to traffic safety study

was provided by Lord (2006). In his paper, Lord (2006) investigated the effect of small sample size and low sample mean problem on estimating the fixed dispersion parameter of negative binomial models. He suggested that in most cases, a sample size of less than 200 can result in an inaccurate estimation of the fixed dispersion parameter. As discussed before, the estimated parameters for OP models using the MLE method based on data of small sample size may also be unreliable. When the BOP is used, by properly choosing prior distributions, the Bayesian method can conceptually produce reliable estimations even when the sample size is small. To validate the advantage of the BOP model with small samples, a sample of 100 records was drawn from the original data. For this small sample, the numbers of no injury, possible injury, nonincapacitated injury, incapacitated injury, and fatal injury crashes are 66, 12, 12, 9, and 1, respectively.

Both OP and BOP models were applied to this small data set. The estimated parameters are shown in Table 4. The prior means and variances were chosen based on experience and the estimated values reported in studies on similar topics (Kockelman and Kweon 2002; Yamamoto and Shankar 2004). The chosen prior means and variances are also listed in Table 4. For ease of description, we refer to this BOP model used here as the fourth BOP model.

Since the original data have larger sample sizes and are expected to better represent the population, the fitted results (in Table 2) from the OP model based on the original data are thus used as the benchmark and compared with the fitted results in Table 4. Comparing Tables 2 and 4, one can see that the estimated means from the fourth BOP model are apparently closer to the corresponding values in Table 2 than those from the OP model. In addition, for the fitted OP model, the coefficients for variables FIRE and ROLLOVER are now either zero or negative, and are

Table 4. Comparison of Parameter Estimation Results between MLE and Bayesian Methods

Parameter	OP		Prior mean	Prior var.	Fourth BOP model	
	Mean	Std.			Mean	Std.
Intercept	-7.1834	0.4046	-0.44	—	-0.3854	0.2473
AGE	-0.0068	0.0087	0.10	—	-0.0069	0.0068
GENDER	-0.3945	0.2704	-0.23	—	-0.2852	0.1998
ALCOHOL	-0.3721	0.9724	0.32	—	0.2975	0.2930
SUV	-0.3749	0.3937	-0.16	—	-0.1439	0.2516
VAN	-0.9513	0.7965	-0.12	—	-0.2265	0.2850
VAGE	0.0009	0.0271	0.10	—	0.0042	0.0222
ROLLOVER	-10.0775	0.0000	1.00	—	0.9051	0.2940
FIRE	0.0000	0.0000	1.00	—	0.9397	0.3011
FRONT	7.2421	0.2344	0.54	—	0.4060	0.2117
RIGHT	7.4583	0.2923	0.30	0.1	0.3307	0.2456
LEFT	7.5260	0.2932	0.50	—	0.5481	0.2515
BACK	7.6897	0.2589	0.25	—	0.4821	0.2330
TOP	0.0000	0.0000	0.30	—	0.2930	0.3151
JUNC	-0.4794	0.2807	-0.20	—	-0.3330	0.2045
ALIGN	0.2777	0.4338	0.10	—	0.1650	0.2466
PROF	0.5167	0.3355	0.10	—	0.3059	0.2312
SURF	-0.1537	0.6220	-0.20	—	-0.0322	0.2233
LIGHT	0.4975	0.3301	0.10	—	0.1959	0.2228
WEATH	0.2806	0.7007	-0.13	—	-0.0012	0.2387
γ_2	0.4140	0.1122	—	—	0.3687	0.0927
γ_3	1.0414	0.1916	—	—	0.9164	0.1643
γ_4	2.1885	0.4007	—	—	2.0932	0.4533
Likelihood ratio		17.8	—	—	—	—
LRI		0.0851	—	—	—	—
Number of observations		100	—	—	—	—
Log-likelihood at zero		-104.6	—	—	—	—
Log-likelihood at convergence		-95.7	—	—	—	—

obviously incorrect. By incorporating the prior knowledge, the BOP model produced reasonable parameters for those two variables. However, note that for variable AGE, the BOP model still could not produce a reasonable parameter even after incorporating the prior information. This is because the data also played an important role in the parameter estimation as shown in Eq. (5), and the parameter estimation was the combined result from the data and the prior information. In this case, even stronger prior information may not solve this discrepancy if the underlying characteristic of the smaller data set is significantly different from that of the larger data set.

A prediction accuracy comparison between the OP and the BOP models was also conducted. The fitted OP and two BOP models (the second and the fourth BOP models) based on the previously discussed small sample data were applied to predict the remaining data. To make the comparison more general, in addition to the first small sample dataset, another three small sample datasets were randomly sampled with sizes equal to 100. We applied the same fitting and prediction procedure to all four small samples. Table 5 summarizes the fitting and prediction performances. The comparison is based on the measures of effectiveness of bias, RMSE, and accuracy as defined in Eqs. (7)–(9)

$$\text{bias: } \frac{1}{N} \sum_{i=1}^N |p_i - o_i| \quad (7)$$

$$\text{RMSE: } \sqrt{\frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2} \quad (8)$$

$$\text{Accuracy: } \frac{\# \text{ of corrected predictions}}{N} \quad (9)$$

where N =total number of fitting or testing observations; p_i =predicted outcome; and o_i =observed outcome.

It can be seen from Table 5 that in some cases, the two BOP models slightly underperform the OP model in terms of fitting bias, RMSE, and accuracies. However, both BOP models considerably outperform the OP model for all prediction bias, RMSE, and accuracies. This may suggest that the information contained in the small sample datasets cannot adequately represent the large dataset or the population. Models fitted completely depending on the small sample data will have poor generalization ability. In this case, incorporation of proper prior information can be an effective means to improve models' generalization abilities. It is interesting to see that even though the second BOP model used normal priors with zero mean and a variance value of 0.5, which did not provide very strong prior information as the fourth BOP model did, its prediction performance in all cases is clearly better than the OP model (Table 5). This seems to suggest that even vague prior information can still improve the generalization ability of regression models when the sample sizes are small.

Table 5. Comparison of Fitting and Prediction Performances

Scenario	Fitting			Prediction		
	Bias	RMSE	Accu. (%)	Bias	RMSE	Accu. (%)
(a) Ordered probit model						
1	0.61	1.18	68.0	0.75	1.34	62.9
2	0.62	1.17	65.0	0.77	1.37	61.7
3	0.69	1.18	59.0	0.78	1.35	59.9
4	0.62	1.08	58.0	0.68	1.21	61.0
(b) Second BOP model						
1	0.61	1.18	68.0	0.69	1.27	64.7
2	0.64	1.20	65.0	0.72	1.31	63.4
3	0.70	1.21	60.0	0.71	1.28	62.7
4	0.65	1.08	55.0	0.67	1.21	61.8
(c) Fourth BOP model						
1	0.64	1.22	67.0	0.65	1.23	65.6
2	0.70	1.30	64.0	0.66	1.24	65.4
3	0.69	1.18	59.0	0.66	1.22	64.6
4	0.67	1.12	55.0	0.65	1.21	64.7

Conclusions and Discussion

This research investigated the application of BOP models in traffic crash injury severity analysis. First, the BOP model was introduced and compared with the commonly used OP model from a theoretical perspective. The BOP model and the classic OP model were then applied to analyze driver's injury severity. A sample of 76,994 records from the 2003 NASSGES was used in this study. The model fitting results from both the BOP and the classic OP models are very close when noninformative prior distributions were used for the BOP model, and these results are also consistent with the results reported by Kockelman and Kweon (2002), and Yamamoto and Shankar (2004). With more informative prior distributions being used for the BOP model, the model fitting results are still not significantly different from those of the OP model. This is due to the large data size that limited the effects of prior information.

To further investigate the benefits of using BOP models, a small sample of 100 records was randomly drawn from the initial large dataset. Both the OP model and a BOP model were applied to this small sample data. The results show that with proper prior setting, the BOP model produces more reasonable parameter estimations than the OP model with this smaller sample.

Three additional small datasets were also sampled from the original large dataset. The OP and two BOP models were fitted based on these small sample datasets and applied to predict the remaining data. The results show that for all four small samples, the BOP models produce better prediction performances than the OP model. Although promising results have been obtained by comparing the performance of the OP model with that of the BOP model based on small sample datasets, it must be pointed out that for large sample datasets the benefits of using BOP models may not be significant, as large sample datasets usually contain enough information to fit an OP model.

This study has demonstrated the BOP model's potential to produce more reasonable parameter estimations and better predictions than the OP model, especially when the sample size is small. The prior distributions of the BOP model may be subjective in nature, however, it provides the BOP model with more flexibility.

Although it is difficult and subjective to choose the prior distributions, it has not prevented the extensive and successful applications of Bayesian models in many fields in recent years. In this preliminary research, normal prior distributions were used for all parameters, and the prior means and variances were determined empirically based on the results reported in studies on similar topics. Future research should be conducted to evaluate the performances of different prior distributions. In addition, the BOP and the OP models should also be compared using other datasets to obtain a more convincing conclusion on the benefits of using the BOP models.

Acknowledgments

The writers would like to thank the anonymous reviewers for their valuable comments that substantially improve the quality of this paper.

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.
- Abdel-Aty, M., and Abdelwahab, H. T. (2004). "Predicting injury severity levels in traffic crashes: A modeling comparison." *J. Transp. Eng.*, 130(2), 204–210.
- Albert, J. H., and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *J. Am. Stat. Assoc.*, 88(422), 669–679.
- 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.
- Chang, L. Y., and Mannering, F. (1999). "Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents." *Accid. Anal Prev.*, 31(5), 579–592.
- Congdon, P. (2003). *Applied Bayesian modelling*, Wiley, West Sussex, U.K.
- Cowles, M. K. (1996). "Accelerating Monte Carlo chain convergence for

- cumulative-link generalized linear models." *Stat. Comput.*, 6(2), 101–111.
- Dissanayake, S., and Ratnayake, I. (2005). "An investigation on severity of rural highway crashes in Kansas." *Proc., 84th Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, D.C.
- Duncan, C. S., Khattak, A. J., and Council, F. M. (1998). "Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions." *Transportation Research Record*, 1635, Transportation Research Board, Washington, D.C., 63–71.
- Greene, W. H. (2000). *Econometric analysis*, 4th Ed., Prentice-Hall, Upper Saddle River, N.J.
- Johnson, V. E., and Albert, J. H. (1999). *Ordinal data modeling*, Springer, New York.
- Khattak, A. J., Pawlovich, M. D., Souleyrette, R. R., and Hallmark, S. L. (2002). "Factors related to more severe older driver traffic crash injuries." *J. Transp. Eng.*, 128(3), 243–249.
- Khattak, A. J., and Targa, F. (2004). "Injury severity and total harm in truck-involved work zone crashes." *Proc., 83rd Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, D.C.
- Khorashadi, A., Niemeier, D., Shankar, V., and Mannering, F. (2005). "Differences in rural and urban driver-injury severities in accidents involving large-trucks: An exploratory analysis." *Accid. Anal Prev.*, 37(5), 910–921.
- Klop, J. R., and Khattak, A. J. (1999). "Factors influencing bicycle crash severity on two-lane, undivided roadways in North Carolina." *Transportation Research Record*, 1674, Transportation Research Board, Washington, D.C., 78–85.
- Kockelman, K. M., and Kweon, Y. J. (2002). "Driver injury severity: An application of ordered probit models." *Accid. Anal Prev.*, 34(3), 313–321.
- Lee, C., and Abdel-Aty, M. (2005). "Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida." *Accid. Anal Prev.*, 37(4), 775–786.
- Lord, D. (2006). "Modeling motor vehicle crashes using poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter." *Accid. Anal Prev.*, 38(4), 751–766.
- Lord, D., and Bonneson, J. A. (2005). "Calibration of predictive models for estimating safety of ramp design configurations." *Transportation Research Record*, 1908, Transportation Research Board, Washington, D.C., 88–95.
- Mckelvey, R. D., and Zavoina, W. (1975). "A statistical model for the analysis of ordinal level dependent variables." *J. Math. Sociol.*, 4, 103–120.
- Milton, J. C., Shankar, V. N., and Mannering, F. L. (2008). "Highway accident severities and the mixed logit model: An exploratory empirical analysis." *Accid. Anal Prev.*, 40(1), 260–266.
- National Highway Traffic Safety Administration (NHTSA). (2003). *General estimates system coding and editing manual*, Washington, D.C.
- O'Donnell, C. J., and 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.
- O'Hagan, A., and Luce, B. R. (2003). "A primer on bayesian statistics in health economics and outcomes research." (<http://www.medtap.com/Method/Bayesian%20Primer.pdf>) (May 23, 2007).
- Riffat, S. M., and Chor, C. H. (2005). "Analysis of severity of single-vehicle crashes in Singapore." *Proc., 84th Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, D.C.
- SAS Institute Inc. (2004). *Version 9.1.3 of the SAS system for Windows*, Cary, N.C.
- Savolainen, P., and Mannering, F. (2007). "Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes." *Accid. Anal Prev.*, 39(5), 955–963.
- Shankar, V., Mannering, F., and Barfield, W. (1996). "Statistical analysis of accident severity on rural freeways." *Accid. Anal Prev.*, 28(3), 391–401.
- Siddiqui, N. A., Chu, X. H., and Guttenplan, M. (2006). "Crossing locations, light conditions, and pedestrian injury severity." *Proc., 85th Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, D.C.
- Srinivasan, K. K. (2002). "Injury severity analysis with variable and correlated thresholds: An ordered mixed logit (OML) formulation." *Proc., 81st Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, D.C.
- Wang, X. K., and Kockelman, K. M. (2005). "Occupant injury severity using a heteroscedastic ordered logit model: Distinguishing the effects of vehicle weight and type." *Proc., 84th Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, D.C.
- Washington, S., Karlaftis, M. G., and Mannering, F. L. (2003). *Statistical and econometric methods for transportation data analysis*, Chapman & Hall/CRC, Boca Raton, Fla.
- Yamamoto, T., and Shankar, V. N. (2004). "Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects." *Accid. Anal Prev.*, 36(5), 869–876.
- Zajac, S. S., and Ivan, J. N. (2003). "Factors influencing injury severity of motor vehicle-crossing pedestrian crashes in rural Connecticut." *Accid. Anal Prev.*, 35(3), 369–379.