Investigation of Effects of Underreporting Crash Data on Three Commonly Used Traffic Crash Severity Models

Multinomial Logit, Ordered Probit, and Mixed Logit

Fan Ye and Dominique Lord

Although much work has been devoted to developing crash severity models to predict the probabilities of crashes for different severity levels, few studies have considered the underreporting issue in the modeling process. Inferences about a population of interest are biased if crash data are treated as a random sample from the population without consideration of the different unreported rates for each crash severity level. The primary objective of this study was to examine the effects of underreporting for three commonly used traffic crash severity models: multinomial logit (MNL), ordered probit (OP), and mixed logit (ML) models. The objective was accomplished with a Monte Carlo approach that used simulated and observed crash data. The results showed that, to minimize the bias and reduce the variability of a model, fatal crashes should be set as the baseline severity for the MNL and ML models, while for the OP models, the rank for the crash severity should be set from fatal to property damage only in a descending order. None of the three models was immune to this underreporting issue. When full or partial information about the unreported rates for each severity level was known, treatment of the crash data as outcome-based samples in model estimation (through the weighted exogenous sample maximum likelihood estimator) dramatically improved the estimation for all three models as compared with the results produced from the maximum likelihood estimator.

Over many years now, a great deal of work has been devoted to the development and application of statistical models for analyzing motor vehicle crashes. It is generally agreed that these statistical models are classified into two categories: crash count and crash severity models (*1*, *2*). The former (e.g., Poisson and Poisson-gamma models) estimate the probability of observing the number of crashes for different severity levels. Crash severity models (e.g., discrete outcome models such as logit or probit models), by contrast, are intended to estimate the probability that a crash will fall into one of the severity levels on condition that the crash has occurred. Crash count and severity models usually are constructed on the basis of police-reported crash data and are used to investigate crash occurrences related to highway design features, environmental conditions, and traffic flow, among other characteristics. It has been well documented, however, that crashes often go unreported, and particularly those associated with relatively low severity levels (*3*–*5*). This underreporting issue can yield to significant biases when used to predict the probability of crash severity (*3*). Numerous studies have investigated factors that influence the unreported rates for different crash severity levels. Few studies, however, have thoroughly investigated underreporting issues related to crash model development.

The primary objective of this study was to examine the effects of underreporting on three commonly used traffic crash severity models: multinomial logit (MNL), ordered probit (OP), and mixed logit (ML) models. More specifically, this study investigated how each of these models performed for different unreported rates. A secondary objective consisted in quantifying how the outcome-based sampling method, through use of the weighted exogenous sample maximum likelihood estimator (WESMLE), could account for specific underreporting conditions when either full or partial knowledge of severity unreported rates was available. The study objectives were accomplished by taking a Monte Carlo approach with the use of simulated and observed crash data.

This paper is divided into five sections. The second section provides background information about the underreporting issue in crash data and is related to crash severity modeling, as well as to the model estimation methods that can account for underreported data. The third section describes the results for the three models for various unreported rates with the use of simulated data. The fourth section presents the modeling results for the three models with the use of observed crash data. The fifth section summarizes the key findings of this study.

BACKGROUND

This section briefly summarizes the literature on underreporting issues associated with crash data and crash severity modeling, and then presents model estimation methods that can be used to account for underreported crash data.

Underreported Crash Data

About 20 years ago, Hauer and Hakkert pointed out that not all traffic crashes were reportable and not all reportable crashes were in fact reported (*3*). This underreporting can limit the ability to manage road safety, because most of the analyses related to road safety are based on reported crash data. The analysis of underreported crash data leads

Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136. Corresponding author: F. Ye, fanclye77@tamu.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2241, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 51–58.
DOI: 10.3141/2241-06

to a biased estimate when crash prediction models are used, which results in ineffective treatments when the models are applied for such purposes. Once the underreporting issue in crash data was recognized, some researchers began to study this topic in greater depth (*3–15*). These studies revealed that crashes were underreported in all industrialized countries, but the unreported rate was worse in developing countries. The probability of reporting was found to be influenced by the crash severity, age of the victim, role of the victim (e.g., whether the victim was the driver or the passenger), and the number of vehicles involved (*3*).

Underreported data tend to produce biased estimations for crash count models and crash severity models. Underreporting affects crash severity models more critically, however, because the reported rates for various severity categories are different. Crashes with a lower severity such as property damage only (PDO) collisions are more likely to go unreported, which leads to the overrepresentation of crashes of a relatively higher severity and to underrepresentation of crashes with lower severity. It has been widely accepted that fatal crashes have the highest reporting rate and PDO crashes have the lowest reporting rate. After reviewing 18 studies in which researchers examined police, hospital, and insurance sources for common entries, Hauer and Hakkert concluded that the unreported rates were 5% for fatality, 20% for injuries that required hospitalization, and perhaps 50% for all injuries (*3*). In a comprehensive meta-analysis, on the basis of 49 studies in 13 countries, Elvik and Mysen found values equal to 5% for fatal injuries, 30% for serious injuries, 75% for slight injuries, and 90% for very slight injuries (*4*). According to Blincoe et al., up to 25% of all minor injuries and almost 50% of PDO crashes are likely to go unreported, because most drivers do not want to have the police involved (or other authoritative figures) on account of insurance concerns or legal repercussions (*16*).

Only a limited number of studies have investigated the effects of underreporting in both the crash count model (*17, 18*) and crash severity model (*19*). As a result, some new approaches have been proposed to account for underreporting in traditional crash model analyses. The next section discusses previous research on underreporting in crash severity modeling.

Underreporting in Crash Severity Modeling

Inconsistent, unreported rates among different severity levels lead to biased results, which in turn can lead to the overestimation of the probability of high-severity crashes and to the underestimation of low-severity crashes, particularly PDO crashes. In addition, underreporting causes biased parameters, which skew the inferences on the effects of key explanatory variables in prediction models. Thus far, only one study has been found that deals with modeling crash severity and underreported data.

Yamamoto et al. investigated the effects of underreporting on parameter estimation for the OP model and the sequential binary probit model (*19*). In their study, the results indicated that the estimates of the explanatory variables and parameter elasticities of both models could be significantly biased if underreporting was not considered. In addition, the researchers regarded traffic crash data as response-based samples with unknown population shares of the injury severities, and used a pseudo-likelihood function to account for the effects of underreporting on parameter estimation for both models (*20, 21*). The population shares of each severity category were estimated for each model that provided insights on the levels of underreporting for each crash severity level. The validation and efficiency of the methods were not confirmed, however. Meanwhile, because only one set of crash data was applied to the models, no information was attained about the model effects on different combinations of unreported rates for each crash severity category.

Model Estimation Methods for Underreported Crash Data

Crash severity models usually are estimated on the basis of random sampling without consideration of the underreporting in crash data. Because of the unique underreporting characteristics in crash data (i.e., unreported rates are different according to the crash severity category), however, crash data should be treated as outcome- or choice-based samples instead of random samples from the population. Without consideration of the underreporting issue for the model, model estimation results would definitely be biased (*19*).

Although it is rare to treat crash data as outcome-based samples, choice-based samples are commonly used in other areas of research, such as transportation economics. Choice-based samples usually are collected by stratifying the data to obtain better information about alternatives that are infrequently chosen in the population when a random sample is not large enough for effective statistical analysis (*22*). Several methods have been developed by economists since 1977 to handle choice-based samples, as summarized in Ye's dissertation (*23*). Among all methods WESMLE is the most consistent and easy to compute, which makes it the most widely used method, although it is not completely efficient. In the research reported in this paper, WESMLE was used for underreported crash data in three crash severity models.

WESMLE is the maximum of the weighted likelihood function in which the weights depend on both the population share of each severity type (the fraction of each severity category in population) and the sample share of each severity type (the fraction of each severity level in an underreported data set). By weighting the observations appropriately, WESMLE makes the outcome-based samples behave asymptotically, as if they were random samples (*24*).

The log likelihood for a WESMLE, as shown in Equation 1, is equivalent to that of the maximum likelihood estimator (MLE), except that each traffic crash is weighted by the ratio of the actual crash severity's population share Q_i, to the sample share H_i, which is the severity share for the underreported crash data.

$$\text{log likelihood for WESMLE} = \sum_{n=1}^{N} \sum_{i \in C_n} d_{ni} \left(\frac{Q_i}{H_i} \right) \ln P\left(i \mid x_{ni}, \beta_i\right) \qquad (1)$$

where

N = number of recorded crashes;
C_n = set of severity categories for individual crash n, C_n = (K = fatal injury, A = incapacitating injury, B = nonincapacitating injury, C = possible injury, and O = PDO) described in the third section of this paper;
d_{ni} = indicator variable equal to 1, if individual crash n belongs to severity level i, and zero otherwise;
x_{ni} = vector of contributing factors associated with individual crash n at severity category C_n;
β_i = vector of the estimable parameters associated with contributing factors x_{ni}; and
$P(i \mid x_{ni}, \beta_i)$ = probability of severity level that belongs to I, given the contributing factors x_{ni} and estimates β_i.

Different models have different probability functions:

For the MNL model:

$$P(i \mid x_{ni}, \beta_i) = \frac{\exp(\alpha_i + \beta_i X_{ni})}{\sum_{\forall i} \exp(\alpha_i + \beta_i X_{ni})} \quad (2)$$

For the ML model:

$$P(i \mid x_{ni}, \beta_i) = \int \frac{\exp[\alpha_i + \beta_i X_{ni}]}{\sum_{\forall i} \exp(\alpha_i + \beta_i X_{ni})} f(\beta_i \mid \theta) d\beta_i \quad (3)$$

And for the OP model:

$$\begin{cases} P(i=1 \mid x_{ni}, \beta) = \phi(\gamma_1 - \beta X_{ni}) \\ P(i \mid x_{ni}, \beta) = \phi(\gamma_i - \beta X_{ni}) - \phi(\gamma_{i-1} - \beta X_{ni}) \\ P(i=5 \mid x_{ni}, \beta) = 1 - \phi(\gamma_{i-1} - \beta X_{ni}) \end{cases} \quad (4)$$

More details of the model structures and probability functions for all three models can be found in Ye's dissertation (*23*).

ANALYSIS WITH SIMULATED DATA

To study the effects of underreporting on three models and to verify the effectiveness of WESMLE for underreported data, a Monte Carlo approach was developed to examine the underreporting with the use of simulated and observed crash data. By repeating the sampling to produce estimates more clustered around the true values, a Monte Carlo approach was an ideal way to verify the underreporting effects on the three models, because the data were created with the knowledge of true values of estimators and true response functions. In addition, various data with different unreported rates could be created a sufficient number of times. Thus the bias could be evaluated by comparing the model estimation with the true values.

Simulation Design

Because the crash data had five severity categories, the number of parameters to investigate was quite large. The crash data were categorized in this study as K, A, B, C, and O. To simplify the analysis, one covariate randomly generated from the standard normal distribution was introduced for all three models. In addition, five outcomes (denoted as Levels 1 to 5) were used to replicate the five severity categories. In addition, covariates were kept at the same values no matter the crash severity of the target observation, because all variables included in a crash severity model are observation-related variables rather than outcome related (*25*). The parameter values for the three models were chosen on the assumption that the results would not be affected significantly by different parameter values.

For the MNL model, the parameters of the covariate were kept the same with a value equal to 1 for each level, $\beta_i = 1$. The constant parameter α_i was equal to 0, 0.5, 1, and 1.5 for Levels 1 to 4. (Level 5 was the baseline outcome with $\alpha_5 = \beta_5 = 0$.) The independent variable x for each level was drawn from a normal distribution with a mean equal to –2 and a variance equal to 1. The error term for each level was drawn independently from a Type I extreme value distribution by obtaining draws from a uniform random distribution and applying the following transformation $-\ln[-\ln(u)]$, where u was a random number drawn from the uniform distribution between 0 and 1. Thus, the error terms gave the following proportions 5.7%, 9.4%, 15.4%, 25.4%, and 44.1% for Levels 1 to 5, respectively, in the population.

For the OP model, the variable parameter β was equal to 1 for each level, x was drawn from a normal distribution with a mean equal to 2.2 and a variance equal to 1, and threshold variable γ_i was 0, 0.8, 1.5, and 2.4 for Levels 1 to 4 (for keeping the population ratios of each outcome as close as those for the MNL model, respectively). The error term was normally distributed for each level. Thus, the error terms gave the following proportions: 6.0%, 10.1%, 15.0%, 24.6%, and 44.3% for Levels 1 to 5, respectively.

For the ML model, the steps for generating the data set were similar to those used in generating the data set for the MNL model. The only difference was that the independent variable was assumed to have randomness in the parameter for Level 1, which followed a normal distribution (mean = 1, variance = 1). The population ratios for each level were 14.1%, 8.7%, 14.3%, 23.6%, and 39.3% for Levels 1 to 5, respectively.

The data sets generated for three models on the basis of the true parameters were treated as the complete data sets, (i.e., the population). The underreported data sets were replicated by randomly removing some data according to the designed unreported rates. To generate sufficient samples even after the random removal of some data, the original sample size was set to be 50,000. In other words, the complete data sets had 50,000 observations for three models, and all the removed observations were considered to be the unreported ones.

Data sets for each model were repeatedly drawn 100 times for each unreported rate designated according to the designed true parameter values of the model. On the basis of the 100 estimated models, the bias of each parameter was calculated as bias = $E(\hat{\beta}_r) - \beta_{\text{baseline}}$, where r was the number of replications ($r = 100$), and β represented each parameter in the model (both constant parameters and variable parameters). The root mean square error (RMSE) of each parameter in a model was calculated by using the equation $\text{RMSE} = \sqrt{\text{bias}^2 + \text{Var}}$, and the total RMSE of all the variable parameters for each model was used to measure the underreporting effects because it comprised both the bias and variability. To summarize the description above, Figure 1 shows the whole process involved in the Monte Carlo analysis on underreporting for simulated data.

Simulation Results

Scenario 1

Five unreported rates—5%, 10%, 20%, 40%, and 80%—were simulated in each level. The change in bias and variability with the increase of the unreported rates for the three models was examined to verify how the number of unreported observations influenced these two items. (For the complete data sets, on the basis of the designed data for the MNL and OP models, the number of observations for the outcome increased from Levels 1 to 5, while for the ML model the number of observations ranked from low to high: Levels 2, 1, 3, 4, and 5, respectively.) For each underreported data set, WESMLE was used to verify whether it could provide a good model estimation on the basis of the known, unreported rates.

After 100 repetitions, summary statistics, such as mean and standard deviation of each parameter for a model, could be calculated. (Because of space constraints, the results are not included in this paper.) The total RMSEs were compared across different unreported rates for each level and for each model (Table 1).

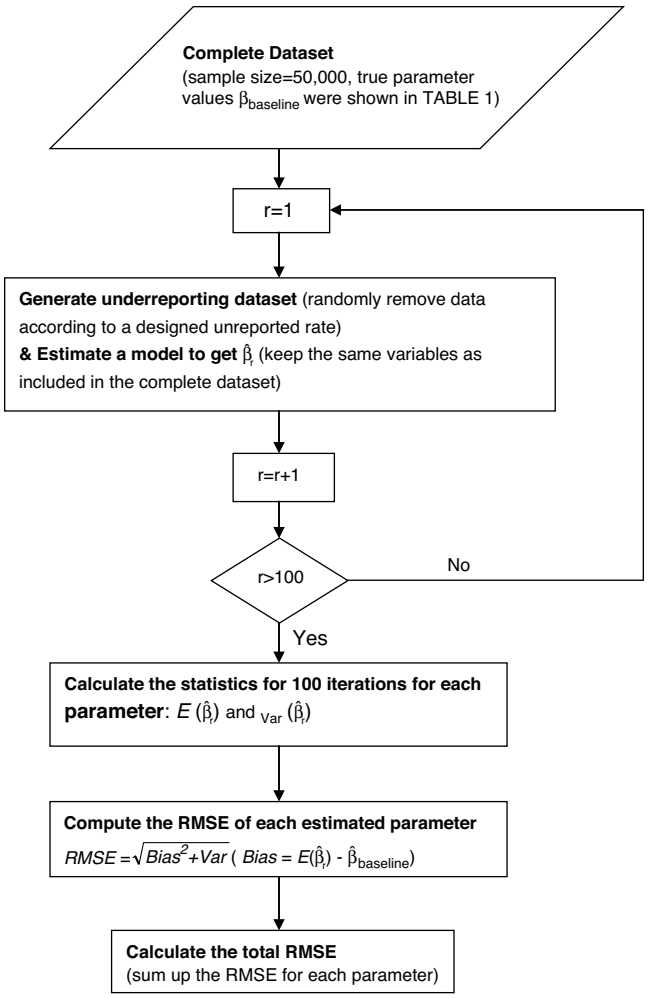

FIGURE 1 Monte Carlo analysis on underreporting for simulated data.

The four key findings for Scenario 1 are as follows:

1. For all three models, with a larger unreported rate, the total RMSE increased with the MLE method. When WESMLE was used, however, to take into account the underreporting issue, and given the variation caused by the randomness in the ML model, the total RMSE remained relatively constant, given the change in the unreported rate.

2. When the MLE was used for model estimation (i.e., without consideration of the underreporting issue in the data), the unreported data showed no clear effects on the total RMSE. Instead, for either the MNL model or the ML model, with the same unreported rate, similar total RMSE values were observed for the parameters from Levels 1 to 4. A much larger value of total RMSE was found when Level 5 contained underreported data. This result was reasonable, because Level 5 was used as the baseline outcome in both the MNL and ML models. The probabilities for the other levels (Levels 1 to 4) were on the basis of the baseline outcome. The underreporting of the baseline outcome would cause more bias in the likelihood function than other levels, and accordingly it would lead to more bias in the model estimation (*23*). This finding indicates that, when the MNL and ML models are used for model estimation with the MLE method, the selection of an outcome with a large, unreported rate as a baseline level should be avoided.

3. The OP model yielded a different result (the largest total RMSE existed when Level 1 was underreported) than the other two models did when outcomes were set up in an ascending order (the outcomes were ranked from Levels 1 to 5). To verify whether the same unreported rate in the level with the lowest order produced the largest total RMSE, the same generated data sets for the OP model were estimated again but in a descending order the second time (from Levels 5 to 1). The total RMSE for each unreported rate is shown in Table 2. From this table, with the same unreported rate when Level 5 was underreported, the total RMSE achieved the largest total, which supported the idea that underreports of the outcome with the lowest rank caused the largest total RMSE. Thus, when the OP model is used for underreported data with the MLE method, the ranking should be avoided of outcomes in an order in which the lowest level has the largest unreported rate.

4. For all three models, the WESMLE method worked well no matter how large the unreported rates and unreported data were for each level. WESMLE gave a more accurate estimation of parameter to the true value; the total RMSE was dramatically lower than the one estimated by the MLE method.

Scenario 2

Although WESMLE performs well for various underreporting situations, the prerequisite for using the method is that the actual, unreported rate is known for each outcome. Usually, it is not fully known for crash data. As shown in Equation 1, WESMLE includes weights in the log likelihood function, which are the ratio of population share Q_i to the sample share H_i for each level. Actually, it is the ratio of the weights, rather than the value of weights themselves, that makes the estimated parameters different and maximizes the log likelihood function of the WESMLE. The ratio of the five weights could be calculated easily as shown below.

Because the weight of level *i* is

$$\text{weight}(i) = \frac{Q_i}{H_i} = \frac{\frac{N_i}{\sum N_i}}{\frac{N_i * (1 - \text{rate}(i))}{\sum [N_i * (1 - \text{rate}(i))]}} \quad (5)$$

where N_i is the number of level *i* in the population, and rate (i) is the unreported rate assumed for level *i*, the ratio of weight (i) for the five levels is

$$\frac{1}{1-\text{rate1}} : \frac{1}{1-\text{rate2}} : \frac{1}{1-\text{rate3}} : \frac{1}{1-\text{rate4}} : \frac{1}{1-\text{rate5}}$$

If full information is available about the unreported rates for all five levels, the above ratio will be the true ratio of weights:

$$\frac{1}{1-\text{Trate1}} : \frac{1}{1-\text{Trate2}} : \frac{1}{1-\text{Trate3}} : \frac{1}{1-\text{Trate4}} : \frac{1}{1-\text{Trate5}}$$

where Trate(i) is the true, unreported rate in level *i*.

Intuitively, the closer the weights ratio is to the true one, the better the estimation will be when WESMLE is used (*23*). To illustrate this idea, a simple example was evaluated. In the simulation, the true, unreported rate was designed to be 40% in one of the five levels. Assume, however, that this number is not known and that the best assumption for it is 20% or 60%. The total RMSE with the use of

TABLE 1 Total RMSE by Unreported Rates with Simulated Data

Outcome in Underreporting	Unreported Rate									
	MLE					WESMLE				
	5%	10%	20%	40%	80%	5%	10%	20%	40%	80%
MNL Model										
Level 1	0.23	0.28	0.39	0.68	1.80	0.21	0.21	0.21	0.22	0.28
Level 2	0.23	0.28	0.39	0.68	1.79	0.21	0.21	0.21	0.22	0.25
Level 3	0.23	0.28	0.39	0.68	1.79	0.20	0.20	0.21	0.21	0.24
Level 4	0.23	0.28	0.40	0.69	1.80	0.21	0.21	0.21	0.21	0.25
Level 5	0.33	0.52	0.99	2.14	6.54	0.21	0.21	0.21	0.23	0.29
OP Model										
Level 1	0.10	0.19	0.39	0.89	2.77	0.06	0.06	0.06	0.06	0.06
Level 2	0.09	0.15	0.28	0.56	1.25	0.06	0.06	0.06	0.06	0.06
Level 3	0.08	0.12	0.21	0.42	0.91	0.06	0.06	0.06	0.06	0.06
Level 4	0.08	0.13	0.23	0.48	1.12	0.06	0.06	0.06	0.06	0.07
Level 5	0.07	0.10	0.17	0.35	1.04	0.06	0.06	0.06	0.06	0.07
ML Model										
Level 1	0.65	0.70	0.85	1.25	2.67	0.64	0.64	0.69	0.76	1.09
Level 2	0.63	0.71	0.79	1.10	2.65	0.60	0.62	0.65	0.65	0.85
Level 3	0.66	0.69	0.78	1.13	2.68	0.64	0.60	0.67	0.66	0.77
Level 4	0.67	0.71	0.88	1.15	2.75	0.64	0.63	0.66	0.66	1.01
Level 5	0.72	0.88	1.29	2.35	7.37	0.65	0.65	0.66	0.69	0.73

these two, unreported rates was calculated for three models, as shown in Table 3. For comparison purposes, the table also lists the estimation results on the basis of the MLE method without taking account of the true, unreported rates.

Table 3 shows that, with the use of WESMLE, the incorrect, unreported rates increased the total RMSE, as compared with instances in which the true underreporting information was used. WESMLE still provided, however, a better estimation than those that did not consider the underreporting in the data (i.e., MLE method). Furthermore, with the use of WESMLE, the incorrect, unreported rates do not refer to any random numbers used as unreported rates. When the assumed, unreported rates shift the weights ratio into another direction (e.g., the weights of five levels are in reverse order to the true ones), the shift might create a larger bias than would use of the MLE method alone. Some sense of the unreported rates for each level is definitely needed to get reasonable results in using WESMLE, even if it is not perfect. In addition, the tentative idea arose that an unreported rate of 20% had a lower total RMSE than the one equal to 60%. Thus, it supports the hypothesis that the closer the weights ratio is to the true value, the better the estimation achieved with the use of WESMLE.

TABLE 2 Total RMSE for OP Model with Outcomes in Descending Order with Simulated Data

Outcome in Underreporting	Unreported Rate				
	5%	10%	20%	40%	80%
MLE					
Level 1	0.07	0.10	0.17	0.37	1.00
Level 2	0.07	0.10	0.17	0.34	0.77
Level 3	0.07	0.10	0.19	0.37	0.81
Level 4	0.10	0.18	0.36	0.75	1.75
Level 5	0.10	0.19	0.37	0.84	2.68
WESMLE					
Level 1	0.06	0.06	0.06	0.06	0.06
Level 2	0.06	0.06	0.06	0.06	0.06
Level 3	0.06	0.06	0.06	0.06	0.06
Level 4	0.06	0.06	0.06	0.06	0.07
Level 5	0.06	0.06	0.06	0.06	0.07

ANALYSIS WITH OBSERVED CRASH DATA

In the previous section of the paper, only one variable was included, which was assumed to be normally distributed. All of the data were generated separately for the three models. Crash severity data involve a large amount of variation, however, which might lead to different patterns of parameter bias and variability. Thus, further analyses were conducted by using observed crash data.

The primary data sources included 4 years (1998–2001) of traffic crash records provided by the Texas Department of Public Safety and the Texas Department of Transportation general road inventory. The crash data focused on single-vehicle crashes that involved fixed objects on rural, two-way highways (crashes that occurred at intersections were excluded). The total of 26,175 usable records in the database contained information related to weather, roadway, driver, and vehicle conditions as well as the severity of the crash reported at the time of the crash (same classification as before). In this data set, there were 11,844 PDO crashes (45.3%), 5,270 Injury C crashes (20.1%), 5,807 Injury B crashes (22.2%), 2,449 Injury A crashes (9.4%), and 805 Fatal crashes (3.1%).

TABLE 3 Total RMSE by Incorrect Unreported Rate with Simulated Data

Outcome in Underreporting	40% (true)		20% (assumed)	60% (assumed)
	MLE	WESMLE	WESMLE	WESMLE
MNL Model				
Level 1	0.68	0.22	0.46	0.59
Level 2	0.68	0.22	0.47	0.60
Level 3	0.68	0.21	0.47	0.62
Level 4	0.69	0.21	0.48	0.62
Level 5	2.14	0.23	1.25	1.69
OP Model				
Level 1	0.89	0.06	0.49	0.70
Level 2	0.56	0.06	0.35	0.54
Level 3	0.42	0.06	0.23	0.34
Level 4	0.48	0.06	0.26	0.37
Level 5	0.35	0.06	0.20	0.29
ML Model				
Level 1	1.25	0.76	0.99	1.15
Level 2	1.10	0.65	0.89	0.99
Level 3	1.13	0.66	0.90	1.04
Level 4	1.15	0.66	0.92	1.06
Level 5	2.35	0.69	1.53	2.16

With the full crash data set, the same three models (MNL, OP, and ML) were developed, and the model estimation from the full data set was considered as the baseline condition for each model. The estimation results from the three models are not included here because of space constraints, but the results can be found in Ye's dissertation (*23*). Next, the underreported crash data sets were generated by randomly removing some crashes for specific severity levels from the full data set according to the designed, unreported rates. For simplicity, 30 underreported data sets were replicated on the basis of a designed, unreported rate in crash data (rather than 100 used for the simulated data). By comparing the results with the baseline conditions, the bias and variance of each parameter were calculated for each model on which the total RMSEs were computed as an index of underreporting. In addition, the same 30 generated, underreported crash data sets for each designed, unreported rate were estimated again for the three models by using the WESMLE.

Similar to Scenario 1 described above, two unreported rates (10% and 40%) were established for each severity level. The unreported crash data sets were applied for three models by using both the MLE and WESMLE methods. The total RMSEs by each unreported rate are shown in Table 4. The OP model was estimated in both ascending and descending order to examine whether the order of severity level had effects on the total RMSE when crash data were underreported.

Table 4 showed that the results were consistent with the simulation output in the previous section. For all three models, the larger unreported rates were associated with a higher total RMSE value. In using WESMLE with the knowledge of the unreported rates, the total RMSE decreased for all underreporting situations for the three models. For the MNL model, when the baseline severity level (fatal or K) was underreported, the total RMSE had a value larger than those attained when other severity levels had the same unreported rate. For the ML model, although the total RMSE value for the PDO underreporting was slightly larger than the baseline severity level (fatal) by the same unreported rate, the value of fatal underreporting was much larger than it was for other severity levels (C, B, A). As mentioned, PDO crashes were more likely to go unreported, whereas fatal crashes usually had the highest reporting rate. Thus, when the MNL and ML models are used to predict the probability of crash severity level, fatal should be set as the baseline outcome to minimize bias and variability. For the OP model, a comparison of the total RMSE values with the use of the MLE method from descending order (KABCO) and ascending order (OCBAK), lower total RMSE values were obtained for the underreporting in O, C, and B when the descending order was used. Because crash data have a more serious underreporting issue for lower severity crashes, use of the descending order provides a better approach to reduce the bias and variability in the estimation of parameters for the OP model.

The analysis described above was done on the basis of only one severity level that was underreported. Further examination was done of the bias and variability of the estimated parameters when different levels of unreported rates were used. The following unreported rates were used: 5%, 20%, 30%, 50%, and 75% for severity KABCO, respectively. The total RMSE values from the MLE and WESMLE methods with the knowledge of real unreported rates are listed in Table 5. As expected, the WESMLE method dramatically decreased the value of total RMSE compared with that of the MLE. The indication is that the WESMLE works well not only when a single crash

TABLE 4 Total RMSE by Different Unreported Rate with Crash Data

Unreported Rate	MNL		ML		OP (KABCO)		OP (OCBAK)	
	MLE	WESMLE	MLE	WESMLE	MLE	WESMLE	MLE	WESMLE
O = 10%	0.37	0.27	1.10	0.98	0.25	0.12	0.33	0.12
C = 10%	0.30	0.23	0.64	0.49	0.11	0.04	0.18	0.04
B = 10%	0.30	0.21	0.76	0.55	0.18	0.08	0.18	0.07
A = 10%	0.34	0.25	0.60	0.48	0.20	0.10	0.18	0.09
K = 10%	0.99	0.90	1.08	1.00	0.24	0.10	0.13	0.08
O = 40%	1.12	0.73	3.37	2.01	1.06	0.32	1.45	0.32
C = 40%	0.92	0.56	1.71	1.08	0.40	0.09	0.70	0.10
B = 40%	0.92	0.53	2.20	1.36	0.67	0.20	0.68	0.17
A = 40%	1.17	0.72	1.71	1.31	0.74	0.22	0.67	0.28
K = 40%	3.01	2.68	3.27	3.01	0.96	0.28	0.46	0.20

TABLE 5 Total RMSE by Unreported Rate for Each Severity Level

Estimation Method	MNL	ML	OP (K–O)	OP (O–K)
MLE	3.84	11.24	2.35	2.56
WESMLE				
Real unreported rates	1.99	6.03	0.68	0.69
Fatal = 5%	4.08	11.50	2.42	2.65
PDO = 50%	3.27	7.65	1.14	1.20

severity is underreported but also when multiple severities have different unreported rates, as long as the unreported rates are known (although this is not always the case for real data).

As discussed in Scenario 2, the change in total RMSE was also examined when partial rather than perfect information was used for the unreported rates. In this case, instead of using 5%, 20%, 30%, 50%, and 75% for severity KABCO for the weight calculation with WESMLE, two hypothetical examples were used. One assumed an unreported rate of 5% in fatal crashes (Example 1), while the other assumed that the unreported rate for the PDO was 50% (Example 2), with all other severity levels kept complete. The results are shown in Table 5. This table illustrates that use of an unreported rate of 50% in PDO crashes decreased the total RMSE more than the MLE for all three models. Use of an unreported rate of 5% in fatal crashes increased the total RMSE. After verifying the ratio of the five severity weights, the above results were found to be reasonable.

The true ratio of weights for KABCO that was used was

$$\frac{1}{1-\text{Trate1}}:\frac{1}{1-\text{Trate2}}:\frac{1}{1-\text{Trate3}}:\frac{1}{1-\text{Trate4}}:\frac{1}{1-\text{Trate5}}=$$
$$\frac{1}{1-5\%}:\frac{1}{1-20\%}:\frac{1}{1-30\%}:\frac{1}{1-50\%}:\frac{1}{1-75\%}$$

For the unreported rate of 5% in fatal crashes in Example 1, the ratio of weights for KABCO was

$$\frac{1}{1-\text{rate1}}:\frac{1}{1-\text{rate2}}:\frac{1}{1-\text{rate3}}:\frac{1}{1-\text{rate4}}:\frac{1}{1-\text{rate5}}=$$
$$\frac{1}{1-5\%}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}$$

For the unreported rate of 50% in PDO crashes in Example 2, the ratio of weights for KABCO was

$$\frac{1}{1-\text{rate1}}:\frac{1}{1-\text{rate2}}:\frac{1}{1-\text{rate3}}:\frac{1}{1-\text{rate4}}:\frac{1}{1-\text{rate5}}=$$
$$\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-0}:\frac{1}{1-50\%}$$

It was obvious that the use of the unreported rate of 5% in fatal crashes shifted the weights ratio into an opposite direction in which the weight of lower crash severities should be larger than the fatal crashes because of the larger unreported rates for the lower crash severity levels. The unreported rate of 50% in PDO crashes, however, still followed the same direction as that of the true weights ratio, in which the weight of PDO was larger than the other severity levels, although it was not as accurate.

The findings further showed what was found in Scenario 2: the closer the weights ratio was to the true one, the better the estimation would be with WESMLE. The incorrect inclusion of the unreported rates in the model estimation for all three models, however, might lead to a worse model estimation with larger bias and variability. Thus, it is important to formulate the weight of each severity level for a model at the same rank as the true one among the five severity levels. Without full knowledge of the true, unreported rates, one conservative way to do so is to include only the unreported rate for the PDO (the largest among all the severity levels) for the weight calculation, with the assumption of a reasonable, unreported rate on the basis of previous research and as much knowledge as possible about the crash data used for estimating the crash severity models.

CONCLUSIONS AND RECOMMENDATIONS

This study aimed to examine the effects of underreporting on three commonly used traffic crash severity models. A secondary objective was to quantify how the outcome-based sampling method in model estimation, through use of the WESMLE, could account for specific underreporting conditions with full and partial knowledge of different severity unreported rates. A Monte Carlo approach with simulated and observed crash data was used to evaluate the three models.

The results of this study showed that the analysis with simulated and observed crash data achieved consistent results on the effects of underreporting for the three models with and without accounting for the underreporting for each crash severity level. To minimize the bias and reduce the variability of the model, fatal crashes should be set as the baseline severity level for the MNL and ML models. For the OP model, the rank of the crash severity should be set from fatal to PDO in a descending order. None of the three models was immune to this underreporting issue.

The results also showed that, when the actual information about the unreported rates of each severity level was known, the WESMLE method dramatically improved the estimation for all three models, compared with the results produced by the MLE, which did not take into account the underreporting issue for crash data. For crash data, however, the unreported rate for each severity level is rarely known with certainty. When partial or imperfect knowledge is available about unreported rates, the WESMLE still gives better estimation results than when no consideration is given of the underreporting in the data (MLE method), although the estimation is not as robust as when the exact underreporting information is obtainable. In addition, the closer the weights ratio is to the true value, the better the estimation will be with the WESMLE. It is hoped that the information provided in this paper will be useful to transportation safety analysts, who are interested in determining factors that influence crash severity.

REFERENCES

1. Lord, D., and F. L. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A*, Vol. 44, No. 5, 2010, pp. 291–305.
2. Savolainen, P. T., F. L. Mannering, D. Lord, and M. A. Quddus. The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. *Accident Analysis and Prevention*, Vol. 43, No. 5, 2011, pp. 1666–1676.
3. Hauer, E., and A. S. Hakkert. Extent and Some Implications of Incomplete Accident Reporting. In *Transportation Research Record 1185*, TRB, National Research Council, Washington, D.C., 1988, pp. 1–10.
4. Elvik, R., and A. B. Mysen. Incomplete Accident Reporting: Meta-Analysis of Studies Made in 13 Countries. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1665, TRB*, National Research Council, Washington, D.C., 1999, pp. 133–140.

5. Tsui, K. L., F. L. So, N. N. Sze, S. C. Wong, and T. F. Leung. Misclassification of Injury Severity Among Road Casualties in Police Reports. *Accident Analysis and Prevention,* Vol. 41, No. 1, 2009, pp. 84–89.
6. Hvoslef, H. *Under-Reporting of Road Traffic Accidents Recorded by the Police at the International Level*. The Norwegian Public Roads Administration, Oslo, 1994.
7. James, J. L., and K. E. Kim. Restraint Use by Children Involved in Crashes in Hawaii, 1986–1991. In *Transportation Research Record 1560*, TRB, National Research Council, Washington, D.C., 1996, pp. 8–12.
8. Stutts, J. C., and W. W. Hunter. Police Reporting of Pedestrians and Bicyclists Treated in Hospital Emergency Rooms. In *Transportation Research Record 1635*, TRB, National Research Council, Washington, D.C., 1998, pp. 88–92.
9. Aptel, I., L. R. Salmi, F. Masson, A. Bourdé, G. Henrion, and P. Erny. Road Accident Statistics: Discrepancies Between Police and Hospital Data in a French Island. *Accident Analysis and Prevention,* Vol. 31, 1999, pp. 101–108.
10. Alsop, J., and J. Langley. Under-Reporting of Motor-Vehicle Traffic Crash Victims in New Zealand. *Accident Analysis and Prevention,* Vol. 33, No. 3, 2001, pp. 353–359.
11. Cryer, P. C., S. Westrup, A. C. Cook, V. Ashwell, P. Bridger, and C. Clarke. Investigation of Bias After Data Linkage of Hospital Admission Data to Police Road Traffic Crash Reports. *Injury Prevention,* Vol. 7, No. 3, 2001, pp. 234–241.
12. Dhillon, P. K., A. S. Lightstone, C. Peek-Asa, and J. F. Kraus. Assessment of Hospital and Police Ascertainment of Automobile Versus Childhood Pedestrian and Bicyclist Collisions. *Accident Analysis and Prevention,* Vol. 33, No. 4, 2001, pp. 529–537.
13. Rosman, D. L. The Western Australian Road Injury Database (1987–1996): Ten Years of Linked Police, Hospital, and Death Records of Road Crashes and Injuries. *Accident Analysis and Prevention,* Vol. 33, No. 1, 2001, pp. 81–88.
14. Amoros, E., J.-L. Martin, and B. Laumon. Under-Reporting of Road Crash Casualties in France. *Accident Analysis and Prevention,* Vol. 38, No. 4, 2006, pp. 627–635.
15. Hauer, E. The Frequency-Severity Indeterminacy. *Accident Analysis and Prevention,* Vol. 38, No. 1, 2006, pp. 78–83.
16. Blincoe, L., A. Seay, E. Zaloshnja, T. Miller, E. Romano, S. Luchter, and R. Spicer. The Economic Impact of Motor Vehicle Crashes, 2000. Publication DOT-HS-809-446. Plans and Policy, NHTSA, U.S. Department of Transportation, 2002.
17. Kumara, S. S. P., and H. C. Chin. Application of Poisson Underreporting Model to Examine Crash Frequencies at Signalized Three-Legged Intersections. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1908,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 46–50.
18. Ma, J. Bayesian Analysis of Underreporting Poisson Regression Model with an Application to Traffic Crashes on Two-Lane Highways. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., 2009.
19. Yamamoto, T., J. Hashijib, and V. N. Shankar. Underreporting in Traffic Accident Data, Bias in Parameters and the Structure of Injury Severity Models. *Accident Analysis and Prevention,* Vol. 40, No. 4, 2008, pp. 1320–1329.
20. Cosslett, S. R. Efficient Estimation of Discrete-Choice Methods. In *Structural Analysis of Discrete Choice Data with Econometric Applications* (C. Manski and D. McFadden, eds.). MIT Press, Cambridge, Mass., 1981, pp. 51–111.
21. Cosslett, S. R. MLE for Choice-Based Samples. *Econometrica,* Vol. 49, 1981, pp. 1289–1316.
22. Bierlaire, M., and D. McFadden. The Estimation of Generalized Extreme Value Models from Choice-Based Samples. *Transportation Research Part B,* Vol. 42, 2008, pp. 381–394.
23. Ye, F. *Investigating the Effects of Underreporting of Crash Data on Three Commonly Used Traffic Crash Severity Models.* PhD dissertation. Texas A&M University, College Station, 2011.
24. Xie, Y., and C. F. Manski. The Logit Model and Response-based Samples. *Sociological Methods and Research,* Vol. 17, No. 3, 1989, pp. 283–302.
25. Khorashadi, A. *Analysis of Driver Injury Severity Logit Models of Truck Involvement/Truck Causation.* PhD dissertation. University of Washington, Seattle, 2003.

The Statistical Methods Committee peer-reviewed this paper.