Investigating the Effects of Underreporting of Crash Data on Three Commonly Used Traffic Crash Severity Models: Multinomial Logit, Ordered Probit and Mixed Logit Models

Fan Ye*
Graduate Research Assistant
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel: (979) 862-8492
Email: fanclye77@tamu.edu

Dominique Lord
Associate Professor
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel. (979) 458-3949
Fax. (979) 845-6481
Email: d-lord@tamu.edu

Dates
July 2010, Revised November 2010, Final Revised March 2011

Total Words
7,462 (5,962 Words, 5 Tables, 1 Figure)

*Corresponding author
ABSTRACT

Although a lot of work has been devoted to developing crash severity models to predict the probabilities of crashes for different severity levels, very few studies have considered the underreporting issue in the modeling process. Inferences about a population of interest will be biased if crash data are treated as a random sample coming from the population without considering the different unreported rates for each crash severity level. The primary objective of this study aimed at examining 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 via a Monte-Carlo approach using simulated and observed crash data. The results showed that in order to minimize the bias and reduce the variability of the 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 (PDO) in a descending order. In addition, none of the three models was immune to this underreporting issue. The results also showed that when the full or partial information about the unreported rates for each severity level is known, treating crash data as outcome-based samples in model estimation, via the Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE), dramatically improve the estimation for all three models compared to the result produced from the Maximum Likelihood estimator (MLE).
1. INTRODUCTION

Over many years now, a lot of work has been devoted on 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 (see 1 and 2 for a thorough review of existing models). 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), on the other hand, are intended to estimate the probability for a crash to fall into one of the severity levels conditional on the fact that the crash has occurred. Crash count and severity models are usually based on police reported crash data and are used for investigating crash occurrences that are related to highway design features, environmental conditions and traffic flow among others. However, it has been well documented that crashes are often unreported, particularly those associated with lower severity levels (3, 4, 5). This underreporting issue can yield to significant biases when used to predict the probability of crash severity (3). There are numerous studies that have investigated factors that influence the unreported rates for different crash severity levels. However, very few studies 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 performs for different unreported rates. A secondary objective consisted in quantifying how the outcome-based sampling method, via the Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE), could account for specific underreporting conditions when the transportation safety analysts had a full or partial knowledge for different severity unreported rates. The study objectives were accomplished via a Monte-Carlo approach using 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 related to crash severity modeling, as well as the model estimation methods that can account for underreported data. The third section describes the results for the three models for various unreported rates using simulated data. The fourth section presents the modeling results for the three models using observed crash data. The fifth section summarizes the key findings of this study.

2. 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.

2.1 Underreported crash data

About twenty years ago, Hauer and Hakkert (3) pointed out that not all traffic crashes were reportable and not all reportable crashes were in fact reported. This can limit the ability to manage road safety, since most of the analyses related to road safety are based on reported crash data. The analysis of underreported crash data would lead to a biased estimate when crash prediction models are used, thus resulting in ineffective treatments when the models are
applied for such purpose. Having realized the underreporting issue in crash data, 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 (whether the victim is the driver, the passenger, or etc.), and number of vehicles involved (3).

Underreported data tend to produce biased estimations for crash count models and crash severity models. However, underreporting is more critical for crash severity models 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 be unreported which leads to the over-representation of crashes with higher severity and under-representation 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 (3) concluded that the unreported rates were 5 percent for fatality, 20 percent for injuries requiring hospitalization and perhaps 50 percent for all injuries. In a comprehensive meta-analysis, based on 49 studies in 13 countries, Elvik and Myssen (4) found values equal to 5 percent for fatal injuries, 30 percent for serious injuries, 75 percent for slight injuries, and 90 percent for very slight injuries. According to Blincoe et al. (16), up to 25 percent of all minor injuries and almost 50 percent of PDO crashes were likely to be non-reported because most drivers did not want to have the police involved (or other authoritative figures) due to insurance concerns or legal repercussions.

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.

2.2 Underreporting in crash severity modeling

The inconsistent unreported rates among different severity levels leads to biased results, which can overestimate the probability of higher severity crashes and underestimate that of lower severity crashes, particularly PDOs. In addition, underreporting causes biased parameters which skew the inferences on the effects of key explanatory variables in prediction models. So far, only one study has been found that deals with modeling crash severity and underreported data.

Yamamoto et al. (19) investigated the effects of underreporting on parameter estimation for the ordered probit model and the sequential binary probit model. 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 (20, 21) to account for the effects of underreporting on parameter estimation for both models. The population shares of each severity category were estimated for each model which provided insights on the levels of underreporting for each crash severity level. However, the validation and efficiency of the methods were not confirmed. Meanwhile, since only one set of crash
data was applied to the models, there was no information attained about the model effects on different combinations of unreported rates for each crash severity category.

### 2.3 Model estimation methods for underreported crash data

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

Though 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 are usually collected by stratifying the data to obtain better information about alternatives that are infrequently chosen in the population when a random sample will not find enough samples for effective statistical analysis (22). There are several methods that have been developed by economists since 1977 to handle choice-based samples, as summarized in Ye’s dissertation (23): Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE), Conditional Maximum Likelihood Estimator, Full information Maximum Likelihood Estimator, Weighted Generalized Method of Moments, Bayesian WESMLE, Smoothed Maximum Likelihood Estimator, and Weighted Conditional Maximum Likelihood Estimator. Among all the methods, though not completely efficient, WESMLE is consistent and easy to compute which makes it the most widely used method. In this research, WESMLE will be used for underreported crash data in three crash severity models.

WESMLE is the maximand of the weighted likelihood function where the weights depend upon 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 dataset). 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.

\[
\text{Log-likelihood for WESMLE} = \sum_{n=1}^{N} \sum_{i \in C_n} d_{ni} \frac{Q_i}{H_i} \ln P(i \mid x_{ni}, \beta_i) \quad (1)
\]

Where,
- \(N\) is the number of recorded crashes;
- \(C_n\) is the set of severity categories from which individual crash \(n\) belongs to, in the study, \(C_n = (K=\text{fatal injury}, A=\text{incapacitating injury}, B=\text{non-incapacitating injury}, C=\text{possible injury}, \text{and } O=\text{PDO})\) which will be described in section 3;
- \(d_{ni}\) is an indicator variable equal to 1 if individual crash \(n\) belongs to severity level \(i\), and 0 otherwise;
\( x_{ni} \) is the vector of contributing factors associated with individual crash \( n \) at severity category \( C_n \);

\( \beta_i \) is a vector of the estimable parameters associated with contributing factors \( x_{ni} \);

\( P(i \mid x_{ni}, \beta_i) \) is the probability of severity level belonging to \( i \) given the contributing factors, \( x_{ni} \), and estimates \( \beta_i \). Different models have different probability functions:

For the MNL model,

\[
P(i \mid x_{ni}, \beta_i) = \frac{\exp(\alpha_i + \beta_i x_{ni})}{\sum_{\forall i} \exp(\alpha_i + \beta_i x_{ni})}
\]

(2)

For the ML model,

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

(3)

For the OP model,

\[
\left\{
\begin{array}{l}
P(i = 1 \mid x_{ni}, \beta) = \phi(\gamma_1 - \beta X_{ni}) \\
P(i \mid x_{ni}, \beta) = \phi(\gamma_1 - \beta X_{ni}) - \phi(\gamma_i - \beta X_{ni}) \\
P(i = 5 \mid x_{ni}, \beta) = 1 - \phi(\gamma_1 - \beta X_{ni})
\end{array}
\right.
\]

(4)

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

3. ANALYSIS USING SIMULATED DATA

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

3.1 Simulation design

Since the crash data have five severity categories, the number of parameters to investigate is very large. The crash data are 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 the same values no matter what crash severity the target observation was since all the variables included in a crash severity model are observation-related rather than outcome-related (25). The parameter values chosen for the three models were based on the assumption that the results would not be affected significantly by different values of the parameters.

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 \( \alpha_i \) was equal to 0, 0.5, 1, 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 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, they 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 \( \beta \) 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 \( \gamma_i \) was 0, 0.8, 1.5, 2.4 for levels 1 to 4 (for keeping the population ratios of each outcome as close as those for the MNL model). The error term was normally distributed for each level. Thus, they 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 dataset were very similar to those used in generating the dataset 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.

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

Datasets for each model were repeatedly drawn 100 times for each unreported rate designated according to the designed true parameter values of the model. Based on the 100 estimated models, the bias of each parameter was calculated as \( Bias = E(\hat{\beta}_i) - \beta_{\text{baseline}} \), where \( r \) was the number of replications (\( r=100 \)), and \( \beta \) 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 using the equation \( RMSE = \sqrt{Bias^2 + Var} \), and the total RMSE of all the variable parameters for each model was used to measure the underreporting effects since it comprised both the bias and variability. As a summary, the whole process described above about the Monte-Carlo analysis on underreporting for simulated data is shown in Figure 1.

### 3.2 Simulation results

#### 3.2.1 Scenario 1

For Scenario 1, the change in bias and variability with the increase of the unreported rates (five unreported rates 5%, 10%, 20%, 40%, and 80% were simulated in each level) for the three models was examined in order to verify how the number of unreported observations influences these two items. (Note: for the complete datasets, based on 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
In addition, for each underreported dataset, WESMLE was used to verify whether it could provide a good model estimation based on the known unreported rates.

After 100 repetitions, summary statistics such as mean and standard deviation of each parameter for a model could be calculated. Due to the space constraint, the results were not included in the paper. In addition, the total RMSEs were compared across different unreported rates for each level and for each model (see Table 1).

There are four key findings for Scenario 1:

1. For all three models, with larger unreported rate, the total RMSE increased using the MLE method. However, when WESMLE is used to take account of the underreporting issue, considering the variation caused by the randomness in the ML model, the total RMSE remained relatively constant given the change in unreported rate.

2. When the MLE was used for model estimation (i.e., without considering the underreporting issue in the data), the unreported data did not show any 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, and while a much larger value of total RMSE was found when level 5 contained underreported data. This is reasonable since level 5 was used as the baseline outcome in both the MNL and ML models. The probabilities for other levels (levels 1 to 4) are based on the baseline outcome, so underreporting of baseline outcome would cause more bias in the likelihood function than other levels and accordingly it leads to more bias in the model estimation (see more explanations in Ye’s dissertation (23)). This indicates that when the MNL and ML models are used for model estimation with the MLE method, the analysts should avoid choosing an outcome with large unreported rate as a baseline level.

3. The OP model has a different result (the largest total RMSE existed when level 1 was underreported) from the other two models when outcomes were setup in an ascending order (the outcomes were ranked from levels 1 to 5). In order to verify whether the same unreported rate in the level with the lowest order produces the largest total RMSE, the same generated datasets for the OP model were estimated again, but in a descending order this time (from levels 5 to 1). The total RMSE for each unreported rate was shown in Table 2. From this table, with the same unreported rate when level 5 was underreported, the total RMSE achieved the largest which supports the idea that underreporting for 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 analysts should avoid ranking the outcomes in an order with the lowest order having the largest unreported rate.

4. The WESMLE method worked well no matter how large the unreported rates and unreported data were for each level for all three models. It gave a more accurate estimation of parameter to the true value with the total RMSE dramatically decreased from that by MLE method.

3.2.2 Scenario 2

Though WESMLE performs well for various underreporting situations, the prerequisite for using the method is that the analysts know the actual unreported rate for each outcome, which usually 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, the ratio of the weights rather than the value of
weights themselves make the estimated parameters different, which maximize log-likelihood function of the WESMLE. The ratio of the five weights could be easily calculated as below.

Since weight of level \( i \) is:

\[
\text{weight } (i) = \frac{Q_i}{H_i} = \frac{\sum N_i}{N_i \ast (1 - \text{rate } (i))} \sum [N_i \ast (1 - \text{rate } (i))]
\]

Where, \( N_i \) is the number of level \( i \) in the population, and \( \text{rate } (i) \) is the unreported rate assumed for level \( i \).

Then the ratio of weight\( (i) \) for the five levels is:

\[
\frac{1}{1-\text{rate } 1} : \frac{1}{1-\text{rate } 2} : \frac{1}{1-\text{rate } 3} : \frac{1}{1-\text{rate } 4} : \frac{1}{1-\text{rate } 5}
\]

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

\[
\frac{1}{1-\text{T rate } 1} : \frac{1}{1-\text{T rate } 2} : \frac{1}{1-\text{T rate } 3} : \frac{1}{1-\text{T rate } 4} : \frac{1}{1-\text{T rate } 5}
\]

Where, \( \text{T rate } (i) \) is the true unreported rate in level \( i \).

Intuitively, the closer the weights ratio is to the true one, better the estimation will be when using WESMLE (23). In order 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, but assume that we do not know this number and our best assumption for it is 20% or 60%. The total RMSE using these two unreported rates were calculated for three models, as shown in Table 3. For comparison, the estimation results based on the MLE method without taking account of the true unreported rates were also listed in this table.

Table 3 showed that the incorrect unreported rates with WESMLE increased the total RMSE compared to those when the true underreporting information was used. However, it still provided a better estimation than those without considering the underreporting in the data (i.e., using the MLE method). Furthermore, the incorrect unreported rates do not refer to any random numbers used as unreported rates with the WESMLE. When the assumed unreported rates shift the weights ratio into another direction (such as making the weights of five levels in a reverse order as the true ones), it might give a larger bias than using the MLE method alone. Some sense of the unreported rates for each level is definitely needed to get reasonable results using WESMLE, even if it is not perfect. In addition, the tentative idea was shown 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 estimation will be using WESMLE.

4. ANALYSIS USING OBSERVED CRASH DATA

In Section 3, we only included one variable which was assumed to be normally distributed. In addition, all the data were generated separately for the three models. However, crash severity data have a large amount of variation which might lead to different patterns of parameter bias and variability. Thus, we conducted further analyses using observed crash data.

The primary data sources included four years (1998-2001) of traffic crash records provided by the Texas Department of Public Safety (TxDPS) and the Texas Department of
Transportation (TxDOT) general road inventory. The crash data focused on single-vehicle crashes involving fixed objects that occurred on rural two-way highways (excluding those occurring at intersections). There were a total of 26,175 usable records in the database which contain 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 dataset, there were 11,844 PDO (45.3%), 5,270 Inj. C (20.1%), 5,807 Inj. B (22.2%), 2,449 Inj. A (9.4%), and 805 Fatal (3.1%) crashes.

First, using the full crash dataset, the same three models (MNL, OP and ML) were developed and the model estimation from the full dataset was considered as the baseline condition for each model. The estimation results from the three models were not included here due to the space constraint, which can be found in the dissertation (23). Next, the underreported crash datasets were generated by randomly removing some crashes for specific severity levels from the full dataset according to the designed unreported rates. For simplicity, 30 underreported datasets were replicated based on 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 upon which the total RMSEs were computed as an index of underreporting. In addition, the same 30 generated underreported crash datasets for each designed unreported rate were estimated again for the three models, using WESMLE.

Similar to Scenario 1 described above, two unreported rates 10% and 40% were established for each severity level. The unreported crash datasets were applied for three models using both the MLE and WESMLE methods. The total RMSEs by each unreported rate were 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. 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 achieved the largest value than the values attained when other severity levels had the same unreported rate. For the ML model, though 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 is much larger than other severity levels (C, B, A). For crash data, as mentioned before, PDO crashes are more likely to be unreported and fatal crash usually has 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 in order to minimize the bias and variability. For the OP model, comparing the total RMSE values using 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. Since crash data have more serious underreporting issue for lower severity crashes, using 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 based on only one severity level that was underreported. Thus, we further examined 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 were listed in Table 5. As expected, the WESMLE method dramatically decreased the value of total RMSE compared to the MLE. It indicates that the WESMLE not only works well when a single crash severity is underreported but also when multiple severities have different unreported rates as long as the unreported rates are known (though not always the case for real data).

As discussed in Scenario 2 above, we also examined the change in total RMSE when we used partial rather than perfect information 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 the unreported rate for the PDO was 50% (example 2), with keeping all other severity levels complete. The results were also shown in Table 5. This table illustrated that using an unreported rate 50% in PDO crashes decreased the total RMSE than that from MLE for all three models, while, using an unreported rate 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

\[
\frac{1}{1 - \text{rate}_1} : \frac{1}{1 - \text{rate}_2} : \frac{1}{1 - \text{rate}_3} : \frac{1}{1 - \text{rate}_4} : \frac{1}{1 - \text{rate}_5} = 1 : 1 : 1 : 1 : 1
\]

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

\[
\frac{1}{1 - \text{rate}_1} : \frac{1}{1 - \text{rate}_2} : \frac{1}{1 - \text{rate}_3} : \frac{1}{1 - \text{rate}_4} : \frac{1}{1 - \text{rate}_5} = 1 : 1 : 1: 1 : 1
\]

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

\[
\frac{1}{1 - \text{rate}_1} : \frac{1}{1 - \text{rate}_2} : \frac{1}{1 - \text{rate}_3} : \frac{1}{1 - \text{rate}_4} : \frac{1}{1 - \text{rate}_5} = 1 : 1 : 1 : 1 : 1
\]

It was obvious that using unreported rate of 5% in fatal crashes shifted the weights ratio into an opposite direction where the weight of lower crash severities should be larger than the fatal crashes due to the larger unreported rates for the lower crash severity levels. However, the unreported rate 50% in PDO crashes 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, though not as accurate.

The findings here further showed what was found in Scenario 2: the closer the weights ratio was to the true one, better the estimation would be with WESMLE. On the other hand, incorrectly including the unreported rates in the model estimation for all three models 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 the same rank as the true one among the five severity levels. Without the full knowledge of the true unreported rates, one conservative way is to only include the unreported rate for the PDO (the largest among all the severity levels) for the weight calculation, assuming a reasonable unreported rate based on previous research and as much knowledge as possible about the crash data used for estimating the crash severity models.
5. CONCLUSIONS AND RECOMMENDATIONS

This paper aimed at studying the effects of underreporting on three commonly used traffic crash severity models. A secondary objective consisted of quantifying how the outcome-based sampling method in model estimation, via WESMLE, can account for specific underreporting conditions when transportation safety analysts have full and partial knowledge for different severity unreported rates. A Monte-Carlo approach using simulated and observed crash data was utilized for evaluating the three models.

The results of this study showed that the analysis using 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. In order 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. It should be pointed out that 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 to the result produced by the MLE which did not take into account the underreporting issue for crash data. However, for crash data, the unreported rate for each severity level is rarely known with certainty. When partial or imperfect knowledge about unreported rates are available, the WESMLE still gives better estimation results than not considering the underreporting in the data (via the MLE method), though 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, better the estimation will be with the WESMLE. It is the hope that the information provided in this paper will be useful for transportation safety analysts who are interested in determining factors that influence crash severity.

REFERENCE


(23) Ye, F. Investigating the Effects of Underreporting of Crash Data on Three Commonly Used Traffic Crash Severity Models. Ph.D. Dissertation. Zachry Department of Civil Engineering, Texas A&M University, College Station, TX, 2011.


LIST OF TABLES AND FIGURES

Tables
TABLE 1 Total RMSE by Different Unreported Rate using Simulated Data
TABLE 2 Total RMSE for the OP model with Outcomes in a Descending Order using Simulated Data
TABLE 3 Total RMSE by Incorrect Unreported Rate using Simulated Data
TABLE 4 Total RMSE by Different Unreported Rates using Crash Data
TABLE 5 Total RMSE by Unreported Rates for Each Severity Level

Figures
FIGURE 1 Monte-Carlo Analysis on Underreporting for Simulated Data.
| Outcome in Underreporting | 5% | 10% | 20% | 40% | 80% | 5% | 10% | 20% | 40% | 80% |
|----------------------------|----|----|----|----|----|----|----|----|----|----|----|
| MLE | 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 |
| WESMLE | 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 |
| OP 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 |
| 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 |
TABLE 2 Total RMSE for the OP model with Outcomes in a Descending Order using Simulated Data

<table>
<thead>
<tr>
<th>Outcome in Underreporting</th>
<th>Unreported Rate</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5%</td>
<td>10%</td>
<td>20%</td>
<td>40%</td>
<td>80%</td>
</tr>
<tr>
<td>MLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>0.07</td>
<td>0.10</td>
<td>0.17</td>
<td>0.37</td>
<td>1.00</td>
</tr>
<tr>
<td>Level 2</td>
<td>0.07</td>
<td>0.10</td>
<td>0.17</td>
<td>0.34</td>
<td>0.77</td>
</tr>
<tr>
<td>Level 3</td>
<td>0.07</td>
<td>0.10</td>
<td>0.19</td>
<td>0.37</td>
<td>0.81</td>
</tr>
<tr>
<td>Level 4</td>
<td>0.10</td>
<td>0.18</td>
<td>0.36</td>
<td>0.75</td>
<td>1.75</td>
</tr>
<tr>
<td>Level 5</td>
<td>0.10</td>
<td>0.19</td>
<td>0.37</td>
<td>0.84</td>
<td>2.68</td>
</tr>
<tr>
<td>WESMLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Level 2</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Level 3</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Level 4</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Level 5</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>Outcome in Underreporting</td>
<td>40% (true)</td>
<td>20% (assumed)</td>
<td>60% (assumed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------</td>
<td>---------------</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>WESMLE</td>
<td>WESMLE</td>
<td>WESMLE</td>
<td></td>
</tr>
<tr>
<td>the MNL model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>0.68</td>
<td>0.22</td>
<td>0.46</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>0.68</td>
<td>0.22</td>
<td>0.47</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>0.68</td>
<td>0.21</td>
<td>0.47</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Level 4</td>
<td>0.69</td>
<td>0.21</td>
<td>0.48</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>Level 5</td>
<td>2.14</td>
<td>0.23</td>
<td>1.25</td>
<td>1.69</td>
<td></td>
</tr>
<tr>
<td>the OP model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>0.89</td>
<td>0.06</td>
<td>0.49</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>0.56</td>
<td>0.06</td>
<td>0.35</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>0.42</td>
<td>0.06</td>
<td>0.23</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Level 4</td>
<td>0.48</td>
<td>0.06</td>
<td>0.26</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>Level 5</td>
<td>0.35</td>
<td>0.06</td>
<td>0.20</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>the ML model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>1.25</td>
<td>0.76</td>
<td>0.99</td>
<td>1.15</td>
<td></td>
</tr>
<tr>
<td>Level 2</td>
<td>1.10</td>
<td>0.65</td>
<td>0.89</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Level 3</td>
<td>1.13</td>
<td>0.66</td>
<td>0.90</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>Level 4</td>
<td>1.15</td>
<td>0.66</td>
<td>0.92</td>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td>Level 5</td>
<td>2.35</td>
<td>0.69</td>
<td>1.53</td>
<td>2.16</td>
<td></td>
</tr>
</tbody>
</table>
# TABLE 4 Total RMSE by Different Unreported Rates using Crash Data

<table>
<thead>
<tr>
<th>unreported rate</th>
<th>MNL</th>
<th>ML</th>
<th>OP(KABCO)</th>
<th>OP(OCBAK)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE  WSMLE</td>
<td>MLE  WSMLE</td>
<td>MLE  WSMLE</td>
<td>MLE  WSMLE</td>
</tr>
<tr>
<td>O=10%</td>
<td>0.37 0.27</td>
<td>1.10 0.98</td>
<td>0.25 0.12</td>
<td>0.33 0.12</td>
</tr>
<tr>
<td>C=10%</td>
<td>0.30 0.23</td>
<td>0.64 0.49</td>
<td>0.11 0.04</td>
<td>0.18 0.04</td>
</tr>
<tr>
<td>B=10%</td>
<td>0.30 0.21</td>
<td>0.76 0.55</td>
<td>0.18 0.08</td>
<td>0.18 0.07</td>
</tr>
<tr>
<td>A=10%</td>
<td>0.34 0.25</td>
<td>0.60 0.48</td>
<td>0.20 0.10</td>
<td>0.18 0.09</td>
</tr>
<tr>
<td>K=10%</td>
<td>0.99 0.90</td>
<td>1.08 1.00</td>
<td>0.24 0.10</td>
<td>0.13 0.08</td>
</tr>
<tr>
<td>O=40%</td>
<td>1.12 0.73</td>
<td>3.37 2.01</td>
<td>1.06 0.32</td>
<td>1.45 0.32</td>
</tr>
<tr>
<td>C=40%</td>
<td>0.92 0.56</td>
<td>1.71 1.08</td>
<td>0.40 0.09</td>
<td>0.70 0.10</td>
</tr>
<tr>
<td>B=40%</td>
<td>0.92 0.53</td>
<td>2.20 1.36</td>
<td>0.67 0.20</td>
<td>0.68 0.17</td>
</tr>
<tr>
<td>A=40%</td>
<td>1.17 0.72</td>
<td>1.71 1.31</td>
<td>0.74 0.22</td>
<td>0.67 0.28</td>
</tr>
<tr>
<td>K=40%</td>
<td>3.01 2.68</td>
<td>3.27 3.01</td>
<td>0.96 0.28</td>
<td>0.46 0.20</td>
</tr>
<tr>
<td>Estimation method</td>
<td>MNL</td>
<td>ML</td>
<td>OP(K-O)</td>
<td>OP(O-K)</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----</td>
<td>-----</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>MLE</td>
<td>3.84</td>
<td>11.24</td>
<td>2.35</td>
<td>2.56</td>
</tr>
<tr>
<td>WESMLE real unreported rates</td>
<td>1.99</td>
<td>6.03</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>fatal=5%</td>
<td>4.08</td>
<td>11.50</td>
<td>2.42</td>
<td>2.65</td>
</tr>
<tr>
<td>PDO=50%</td>
<td>3.27</td>
<td>7.65</td>
<td>1.14</td>
<td>1.20</td>
</tr>
</tbody>
</table>
Complete Dataset
(sample size=50,000, true parameter values $\beta_{\text{baseline}}$ were shown in TABLE 1)

Generate underreporting dataset (Randomly remove data according to a designed unreported rate)
& Estimate a model to get $\hat{\beta}_i$ (keep the same variables as included in the complete dataset)

Calculate the statistics for 100 iterations for each parameter: $E(\hat{\beta}_i)$ and $\text{Var}(\hat{\beta}_i)$

Compute the RMSE of each estimated parameter
$RMSE = \sqrt{\text{Bias}^2 + \text{Var}(\text{Bias} = E(\hat{\beta}_i) - \beta_{\text{baseline}})}$

Calculate the total RMSE
(sum up the RMSE for each parameter)

FIGURE 1 Monte-Carlo Analysis on Underreporting for Simulated Data.