Accident Analysis and Prevention 40 (2008) 1320–1329

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Underreporting in traffic accident data, bias in parameters and the structure of injury severity models

Toshiyuki Yamamoto[a,*], Junpei Hashiji[b,1], Venkataraman N. Shankar[c,2]

[a] Department of Civil Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan
[b] Chodai Co., Ltd., 1-20-4 Nihonbashi-Kakigara-cho, Chuo-ku, Tokyo 103-0014, Japan
[c] Department of Civil and Environmental Engineering, Pennsylvania State University, 226C Sackett Building, University Park, PA 16802, United States

ARTICLE INFO

Article history:
Received 31 October 2005
Received in revised form 17 October 2007
Accepted 29 October 2007

Keywords:
Injury severity
Ordered-response probit model
Sequential probit model
Underreporting

ABSTRACT

Injury severities in traffic accidents are usually recorded on ordinal scales, and statistical models have been applied to investigate the effects of driver factors, vehicle characteristics, road geometrics and environmental conditions on injury severity. The unknown parameters in the models are in general estimated assuming random sampling from the population. Traffic accident data however suffer from underreporting effects, especially for lower injury severities. As a result, traffic accident data can be regarded as outcome-based samples with unknown population shares of the injury severities. An outcome-based sample is overrepresented by accidents of higher severities. As a result, outcome-based samples result in biased parameters which skew our inferences on the effect of key safety variables such as safety belt usage. The pseudo-likelihood function for the case with unknown population shares, which is the same as the conditional maximum likelihood for the case with known population shares, is applied in this study to examine the effects of severity underreporting on the parameter estimates. Sequential binary probit models and ordered-response probit models of injury severity are developed and compared in this study. Sequential binary probit models assume that the factors determining the severity change according to the level of the severity itself, while ordered-response probit models assume that the same factors correlate across all levels of severity. Estimation results suggest that the sequential binary probit models outperform the ordered-response probit models, and that the coefficient estimates for lap and shoulder belt use are biased if underreporting is not considered. Mean parameter bias due to underreporting can be significant. The findings show that underreporting on the outcome dimension may induce bias in inferences on a variety of factors. In particular, if underreporting is not accounted for, the marginal impacts of a variety of factors appear to be overestimated. Fixed objects and environmental conditions are overestimated in their impact on injury severity, as is the effect of separate lap and shoulder belt use. Combined lap and shoulder belt usage appears to be unaffected. The parameter bias is most pronounced when underreporting of possible injury accidents in addition to property damage only accidents is taken into account.

1. Introduction

Injury severities in traffic accidents are usually recorded on ordinal scales (for example, the national Fatal Accident Reporting System). Five categories are typically used: property damage only, possible injury, evident injury, disabling injury and fatality. Disaggregate statistical models which treat such severity categories as ordered or discrete outcomes are plausible methods. The severity outcome is the dependent variable in such methods and driver factors, vehicle characteristics, road and environmental conditions at the time of the accident are extracted from individual accident reports as potential independent variables correlated with injury severity. The parameters (coefficients) in the models are estimated usually assuming random sampling of severities from the population. However, accident data samples cannot be regarded as random samples. Lower severity outcomes are more likely to be under-reported and as a result, accident samples are usually over-represented by injury outcomes of greater severities. The level of underreporting decreases with increased severity. Hauer and Hakkert (1988) and Elvik and Mysen (1999) carried out meta-analyses on incomplete accident reporting. The results of both studies suggest that lower severity accidents suffer from lower reporting rates in general. The net effect of underreporting is outcome-based sampling effects in accident data. In outcome-based sam-

[*] Corresponding author. Tel.: +81 52 789 4636; fax: +81 52 789 3565.
E-mail addresses: yamamoto@civil.nagoya-u.ac.jp (T. Yamamoto), hashiji-j@chodai.co.jp (J. Hashiji), shankarv@engr.psu.edu (V.N. Shankar).
[1] Tel.: +81 3 3639 3301.
[2] Tel.: +1 814 865 9434; fax: +1 814 863 7304.

0001-4575/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2007.10.016

pling of accident data, unique statistical problems arise. First and foremost is the problem of unknown population shares of severity. As a result, estimation of parameters in outcome-based samples with unknown population shares becomes a main challenge.

Consistent estimators have been developed for outcome-based sampling data with unknown shares, but the estimators have not been applied to accident data yet. Thus, the main objectives of this paper are to examine the effects of underreporting on the parameter estimates, and resulting biased estimates of impact of explanatory variables on the injury severity. This has significant implications for inference on the marginal effect of safety interventions such as seat belts. Since seat belt usage is likely to suffer underreporting similar to severity outcome, a method to estimate the unbiased parametric effect of seat belt usage is useful. The rest of this paper presents in order, literature review, methodology, data, results and summary and conclusions from the empirical analysis explicitly accounting for the outcome-based sampling scheme for ordered-response and binary outcome models of injury severity.

2. Literature review

Consistent estimators have been developed and widely applied to outcome-based sampling data in the transportation research field. The main application has however occurred in the area of discrete choice involving for example, choice of mode or vehicle type. Due to the fact these estimators have been applied to the context of "choice of an outcome," they are popularly referred to as choice-based sampling estimators. The weighted exogenous sampling maximum likelihood (WESML) estimator is one of the most popular estimators and often applied to take into account choice-based samples with known population shares. On the other hand, consistent estimators for choice-based samples with unknown sampling rate are not popular, at least in the transportation research field. Cosslett (1981a,b) suggested a pseudo-likelihood function applicable for choice-based samples with unknown population shares. Cosslett also showed in the numerical exercise that the parameter estimates can be biased if the conventional maximum likelihood estimator (MLE) is applied for choice-based samples with unknown population shares, and that the biases can be prevented by applying the suggested pseudo-likelihood function. The pseudo-likelihood function suggested by Cosslett is applied in this study to examine the effects of underreporting on parameter estimates for traffic accident data. The estimator also provides the estimates of the population shares of each category. The estimated population shares provide insights on the levels of underreporting in each injury severity as well. Recently more efficient estimators have been proposed using smoothed likelihood functions (Cosslett, 2005), but this application remains as a future task.

Several popular injury severity model structures have been studied in the extant literature. Sequential binary logit models (e.g., Saccomanno et al., 1996; Dissanayake and Lu, 2002a,b), ordered-response probit models (e.g., O'Donnell and Connor, 1996; Duncan et al., 1998; Khattak, 2001; Kockelman and Kweon, 2002; Abdel-Aty and Keller, 2005; Shimamura et al., 2005), and multinomial or nested logit models (e.g., Shankar and Mannering, 1996; Shankar et al., 1996, 2000; Chang and Mannering, 1999; Carson and Mannering, 2001; Lee and Mannering, 2002; Ulfarsson and Mannering, 2004) are typical examples. The sequential binary logit and ordered-response probit models represent the ordinality in the discrete categories of the injury severity. However, the two models are structurally different. Sequential binary logit models assume that the factors determining the level of the severity change according the level of the severity itself, while ordered-response probit models assume that the same factors are correlated with all levels of injury severity.

On the other hand, ordinality is not theoretically implemented in multinomial and nested logit models; thus information relating to ordering of severities is not inherently captured in those structures. However, the multinomial and nested logit models are structurally flexible in the sense that independent variables are not forced to be the same across all severities. This flexibility is useful from an engineering standpoint, for it allows different sets of independent variables to be identified as significantly associated with different severities. For this reason, multinomial and nested logit models have been applied and succeeded in finding the influencing factors in empirical severity analyses. Another advantage of multinomial logit models is the consistency of the coefficient estimates except constant terms even when the conventional MLE is applied to the outcome-based samples (Cosslett, 1981a,b). Thus, the unknown parameters except constants can be consistently estimated by the conventional MLE for the injury severity data with underreporting if multinomial logit models are applied.

Sequential binary logit models are the combination of several binary logit models, so the parameter estimates except constant terms are not biased, behaving similarly to multinomial logit models. One of the constraints in sequential binary logit models is the assumption of the independence among the error terms of each level of the injury severity as pointed out by McCarthy and Madanat (1994). A significant finding from the nested logit structures (see for example Shankar et al., 1996) is that error correlations may be significant among severities. To address this issue as well, correlations between successive levels of the injury severities are investigated by developing bivariate binary probit models in this study. The assumption of the normal distribution for the error terms enables an appropriate treatment of the correlations. The consistency of the parameter estimates except constant does not hold theoretically when the probit models are used instead of the logit models. However, the difference between the two models is only in the assumption of the error distribution: the normal distribution and the Gumbel distribution. Thus, the bias on the parameter estimates except constant might be minimal, and should be investigated empirically.

Ordered-response probit models may provide biased estimates if the conventional MLE is applied for injury severity data with underreporting. In this study, the ordered-response probit models are applied to examine the effects of the underreporting on the parameter estimates. In addition, sequential binary probit models are applied and compared with the ordered-response probit models. The use of sequential probit models, however, may result in biased coefficient estimates, thus is empirically examined in this study.

Based on the synthesis of the extant literature, we establish two important issues related to unbiased estimation of severity parameters while using sequential or ordered-response structures. The first issue relates to the assumption of known population shares of severity. This is not established as a set of point values for the various severity levels in the literature, so unknown population shares of injury severity are a more reasonable assumption. Given this assumption, the second important issue this paper addresses is the use of pseudo-likelihood function for estimation. While sequential binary outcome models are unbiased in outcome-based samples with known population shares, they are not necessarily unbiased unless a pseudo-likelihood function is employed for samples with unknown population injury shares.

3. Methodology

An ordered-response probit model supposes an unobserved latent and continuous injury severity given as

$$y_i^* = \boldsymbol{\beta}\mathbf{x}_i + \varepsilon_i, \tag{1}$$

where y_i^* is an unobserved latent and continuous injury severity function for accident i, $\boldsymbol{\beta}$ the vector of parameters to be estimated; and $\mathbf{x}_i$ and ε_i are vectors of explanatory variables and a normally distributed random error term, respectively. Injury severity is observed as an ordered category, y_i, given as

$$\begin{aligned} y_i \ &= 1 \quad \text{if}\, y_i^* \le \mu_1, \\ &= 2 \quad \text{if}\, \mu_1 < y_i^* \le \mu_2, \\ &= 3 \quad \text{if}\, \mu_2 < y_i^* \le \mu_3, \\ &\ \vdots \\ &= J \quad \text{if}\, \mu_{J-1} < y_i^*, \end{aligned} \tag{2}$$

where the μ's are unknown parameters to be estimated and ordered from the lowest severity to the highest; and J is the number of severity outcome categories. It should be noted that μ_1 is normalized to 0 without loss of generality. Assuming that ε_i is normally distributed, and normalizing the mean and variance of ε_i to 0 and 1, respectively without loss of generality, we have the following probability:

$$\Pr(y_i = j|\mathbf{x}_i) = \int_{\mu_{j-1}-\boldsymbol{\beta}\mathbf{x}_i}^{\mu_j-\boldsymbol{\beta}\mathbf{x}_i} \phi(\varepsilon_i)\,\mathrm{d}\varepsilon_i = \Phi(\mu_j - \boldsymbol{\beta}\mathbf{x}_i) - \Phi(\mu_{j-1} - \boldsymbol{\beta}\mathbf{x}_i), \tag{3}$$

where ϕ and Φ are the standard normal probability density and cumulative distribution functions, respectively, and μ_0, and μ_J are $-\infty$ and ∞, respectively. Under random sampling assumptions, the unknown parameters can be estimated by a conventional MLE which is given as

$$\begin{aligned} \max \ln L &= \sum_{i=1}^{N} \ln \Pr(y_i = j_i|\mathbf{x}_i) \\ &= \sum_{j=1}^{J} \sum_{i\in(j_i=j)} \ln\{\Phi(\mu_j - \boldsymbol{\beta}\mathbf{x}_i) - \Phi(\mu_{j-1} - \boldsymbol{\beta}\mathbf{x}_i)\}, \end{aligned} \tag{4}$$

where N is the sample size.

On the other hand, a sequential binary probit model supposes that the factors determining the severity outcome change according the level of the severity itself. Sequencing of the binary probit across the range of severities can be done in two ways. The first approach is from the lowest severity to the highest severity, and the second from the highest severity to the lowest. In the first approach, whether the injury severity of accident is observed as property damage only or higher is determined by a vector of explanatory variables at first. Then for the cases with higher injury severity, whether the injury severity is observed as the least severe injury or higher is determined by another vector of explanatory variables. This process continues to the highest severity. The sequential structure of the model represents ordinality of the underlying process. On the other hand in the second approach, whether the injury severity is observed as the highest severity or lower is determined at first. Thus, the two approaches provide different sets of parameter estimates and goodness-of-fit statistics from each other even using the same data set. It is because the two approaches are based on different assumptions on underlying mechanisms of the effects of the explanatory variables. The literature has not suggested a clear superiority between the two approaches, so the first approach is used in this study from the view point mentioned below. A sequential binary probit model supposes J latent variables given as

$$\begin{aligned} y_{1i}^* &= \boldsymbol{\beta}_1\mathbf{x}_i + \varepsilon_{1i}, \\ y_{2i}^* &= \boldsymbol{\beta}_2\mathbf{x}_i + \varepsilon_{2i}, \\ &\ \vdots \\ y_{J-1,i}^* &= \boldsymbol{\beta}_{J-1}\mathbf{x}_i + \varepsilon_{J-1,i}, \end{aligned} \tag{5}$$

where y_{ji}^*'s are unobserved latent variables determining whether the injury severity of accident i is observed as j or higher than j, $\boldsymbol{\beta}_j$'s the vectors of parameters to be estimated; and ε_{ji}'s are random error terms. Injury severity is observed as an ordered category, y_i, given as

$$\begin{aligned} y_i \ &= 1 \quad \text{if}\, y_{1i}^* \le 0, \\ &= 2 \quad \text{if}\, 0 < y_{1i}^* \,\text{and}\, y_{2i}^* \le 0, \\ &= 3 \quad \text{if}\, 0 < y_{1i}^*,\ 0 < y_{2i}^* \,\text{and}\, y_{3i}^* \le 0, \\ &\ \vdots \\ &= J \quad \text{if}\, 0 < y_{1i}^*,\ 0 < y_{2i}^*, \ldots, \text{ and } 0 < y_{J-1,i}^*. \end{aligned} \tag{6}$$

Assuming the independence among the error terms of each level, the probability that accident i is observed is written as

$$\begin{aligned} &\Pr(y_i = 1|\mathbf{x}_i) = \Phi(-\boldsymbol{\beta}_1\mathbf{x}_i), \\ &\Pr(y_i = 2|\mathbf{x}_i) = \Phi(\boldsymbol{\beta}_1\mathbf{x}_i)\Phi(-\boldsymbol{\beta}_2\mathbf{x}_i), \\ &\Pr(y_i = 3|\mathbf{x}_i) = \Phi(\boldsymbol{\beta}_1\mathbf{x}_i)\Phi(\boldsymbol{\beta}_2\mathbf{x}_i)\Phi(-\boldsymbol{\beta}_3\mathbf{x}_i), \\ &\ \vdots \\ &\Pr(y_i = J|\mathbf{x}_i) = \prod_{j=1}^{J-1}\Phi(\boldsymbol{\beta}_j\mathbf{x}_i), \end{aligned} \tag{7}$$

and assuming random sampling, the MLE is given as

$$\begin{aligned} \max \ln L &= \sum_{i=1}^{N} \ln \Pr(y_i = j_i|\mathbf{x}_i) \\ &= \sum_{i\in(j_i=1)} \ln[\Phi(-\boldsymbol{\beta}_1\mathbf{x}_i)] + \sum_{i\in(j_i=2)} \ln[\Phi(\boldsymbol{\beta}_1\mathbf{x}_i)\Phi(-\boldsymbol{\beta}_2\mathbf{x}_i)] \\ &\quad + \sum_{i\in(j_i=3)} \ln[\Phi(\boldsymbol{\beta}_1\mathbf{x}_i)\Phi(\boldsymbol{\beta}_2\mathbf{x}_i)\Phi(-\boldsymbol{\beta}_3\mathbf{x}_i)] + \cdots + \sum_{i\in(j_i=J)} \ln\left[\prod_{j=1}^{J-1}\Phi(\boldsymbol{\beta}_j\mathbf{x}_i)\right] \\ &= \sum_{i\in(j_i=1)} \ln[\Phi(-\boldsymbol{\beta}_1\mathbf{x}_i)] + \sum_{i\in(j_i\ge 2)} \ln[\Phi(\boldsymbol{\beta}_1\mathbf{x}_i)] \\ &\quad + \sum_{i\in(j_i=2)} \ln[\Phi(-\boldsymbol{\beta}_2\mathbf{x}_i)] + \sum_{i\in(j_i\ge 3)} \ln[\Phi(\boldsymbol{\beta}_2\mathbf{x}_i)] \\ &\ \vdots \\ &\quad + \sum_{i\in(j_i=J-1)} \ln[\Phi(-\boldsymbol{\beta}_{J-1}\mathbf{x}_i)] + \sum_{i\in(j_i=J)} \ln[\Phi(\boldsymbol{\beta}_{J-1}\mathbf{x}_i)]. \end{aligned} \tag{8}$$

The equation suggests that the binary probit model determining each level of severity can be estimated separately if the independence among the error terms of each level is assumed. Moreover, the ability of separate estimation is useful if the possibility of underreporting is taken into account. The underreporting is more probable in lower severity levels, so the parameters related to the higher severity levels can be more likely estimated without underreporting biases if the sequence from the lowest to the highest is applied, and the biases from underreporting can be limited to the parameters related to the lower severity levels. For example in Eq. (8), if the sample cases of the lowest severity (j = 1 in Eq. (8)) have underreporting, but the other accidents of higher severities have no underreporting, $\boldsymbol{\beta}_2$ to $\boldsymbol{\beta}_{J-1}$ in Eq. (8) can be separately and consistently estimated by the conventional MLE because the sample cases of the lowest severity are included in the first term alone on the right-hand-side and not in the other terms of Eq. (8). This is the reason why the sequence from the lowest to the highest is applied in this study.

The independence among the error terms of each level is assumed so far, but as McCarthy and Madanat (1994) pointed out, the error terms may be correlated. This is the reason why McCarthy and Madanat proposed the ordered-response logit model. However, the correlation can be also implemented in the sequential binary probit model. The correlations are thought to be stronger between closer levels of the severity if correlations exist. In this study, the correlations between successive two levels of the severity are investigated in order; that is, "first-order" correlations are examined. When ε_{ki} and $\varepsilon_{(k+1)i}$ are assumed to be correlated and the other error terms are assumed independent, the joint probability for $0 < y^*_{ki}$ and $y^*_{(k+1)i} \leq 0$ and that for $0 < y^*_{ki}$ and $0 < y^*_{(k+1)i}$ become different from those used in Eq. (7). The joint probabilities are given as

$$\begin{aligned}
&\Pr(0 < y^*_{ki}, y^*_{(k+1)i} \leq 0|\mathbf{x}_i) = \Pr(0 < \boldsymbol{\beta}_k\mathbf{x}_i + \varepsilon_{1i}, \boldsymbol{\beta}_{k+1}\mathbf{x}_i + \varepsilon_{2i} \leq 0)\\
&\quad = \Pr(-\varepsilon_{1i} < \boldsymbol{\beta}_k\mathbf{x}_i, \varepsilon_{2i} \leq -\boldsymbol{\beta}_{k+1}\mathbf{x}_i) = \Phi_2(\boldsymbol{\beta}_k\mathbf{x}_i, -\boldsymbol{\beta}_{k+1}\mathbf{x}_i, -\rho_{k,k+1}),\\
&\Pr(0 < y^*_{ki}, 0 < y^*_{(k+1)i}|\mathbf{x}_i) = \Pr(0 < \boldsymbol{\beta}_k\mathbf{x}_i + \varepsilon_{1i}, 0 < \boldsymbol{\beta}_{k+1}\mathbf{x}_i + \varepsilon_{2i})\\
&\quad = \Pr(-\varepsilon_{1i} < \boldsymbol{\beta}_k\mathbf{x}_i, -\varepsilon_{2i} \leq \boldsymbol{\beta}_{k+1}\mathbf{x}_i) = \Phi_2(\boldsymbol{\beta}_k\mathbf{x}_i, \boldsymbol{\beta}_{k+1}\mathbf{x}_i, \rho_{k,k+1}),
\end{aligned} \tag{9}$$

where Φ_2 is the standard bivariate normal cumulative distribution function, and $\rho_{k,\,k+1}$ is the correlation between ε_{ki} and $\varepsilon_{k+1,\,i}$. By replacing appropriate parts of Eq. (7) by Eq. (9), Eq. (8) is rewritten as

$$\begin{aligned}
\max \ln L \;=\; & \sum_{i\in(j_i=1)} \ln[\Phi(-\boldsymbol{\beta}_1\mathbf{x}_i)] + \sum_{i\in(j_i\geq 2)} \ln[\Phi(\boldsymbol{\beta}_1\mathbf{x}_i)]\\
& + \sum_{i\in(j_i=2)} \ln[\Phi(-\boldsymbol{\beta}_2\mathbf{x}_i)] + \sum_{i\in(j_i\geq 3)} \ln[\Phi(\boldsymbol{\beta}_2\mathbf{x}_i)]\\
& \vdots\\
& + \sum_{i\in(j_i=k)} \ln[\Phi(-\boldsymbol{\beta}_k\mathbf{x}_i)]\\
& + \sum_{i\in(j_i=k+1)} \ln[\Phi_2(\boldsymbol{\beta}_k\mathbf{x}_i, -\boldsymbol{\beta}_{k+1}\mathbf{x}_i, -\rho_{k,k+1})]\\
& + \sum_{i\in(j_i\geq k+2)} \ln[\Phi_2(\boldsymbol{\beta}_k\mathbf{x}_i, \boldsymbol{\beta}_{k+1}\mathbf{x}_i, \rho_{k,k+1})]\\
& \vdots\\
& + \sum_{i\in(j_i=J-1)} \ln[\Phi(-\boldsymbol{\beta}_{J-1}\mathbf{x}_i)] + \sum_{i\in(j_i=J)} \ln[\Phi(\boldsymbol{\beta}_{J-1}\mathbf{x}_i)].
\end{aligned} \tag{10}$$

As mentioned previously, sample accident records are assumed to be underreported with unknown population shares. Cosslett (1981a,b) suggested a pseudo-likelihood function for outcome-based samples with unknown population shares. According to Cosslett (1993), the log likelihood for stratified sampling can be written as

$$\ln L = \sum_{i=1}^{N} \ln \frac{\Pr(y_i = j_i|\mathbf{x}_i)h(\mathbf{x}_i)}{Q(j_i)}, \tag{11}$$

where $h(\mathbf{x}_i)$ is the density of $\mathbf{x}_i$, and $Q(j)$ is the unknown population share of severity j to be estimated, satisfying $\sum_{j=1}^{J} Q(j) = 1$ and

$$Q(j) = \int h(\mathbf{x}) \Pr(j|\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{12}$$

Thus, $h(\mathbf{x}_i)$ is not independent from y_i, and cannot be ignored when maximizing the function. Cosslett (1981a,b) proposed to replace $h(\mathbf{x})$ by its nonparametric maximum likelihood estimator (NPMLE). The NPMLE of $h(\mathbf{x})$ is the discrete density with weight given as

$$\omega_i = \frac{1}{N\sum_{j=1}^{J}[H(j)/\hat{Q}(j)]\Pr(j|\mathbf{x}_i)}, \tag{13}$$

where the estimated shares $\hat{Q}(j)$ are determined by substituting Eq. (13) into the equations

$$\hat{Q}(j) = \sum_{i=1}^{N} \omega_i \Pr(j|\mathbf{x}_i) \quad \text{for } j = 1, \ldots, J, \tag{14}$$

and $\sum_{i=1}^{N} \omega_i = 1$. The resulting concentrated likelihood is

$$\ln L = \sum_{i=1}^{N} \ln \frac{[H(j_i)/\hat{Q}(j_i)]\Pr(y_i = j_i|\mathbf{x}_i)}{\sum_{k=1}^{J}\{[H(k)/\hat{Q}(k)]\Pr(y_i = k|\mathbf{x}_i)\}}. \tag{15}$$

This can be expressed in a more convenient form given as

$$\ln L = \sum_{i=1}^{N} \ln \frac{[H(j_i)/Q(j_i)]\Pr(y_i = j_i|\mathbf{x}_i)}{\sum_{k=1}^{J}\{[H(k)/Q(k)]\Pr(y_i = k|\mathbf{x}_i)\}}. \tag{16}$$

The first-order conditions for maximization of Eq. (16) over Q become Eq. (14) after substituting for ω_i. Thus, maximization of Eq. (16) over both structural parameters in $\Pr(y_i = |\mathbf{x}_i)$ and Q gives consistent estimators. Eq. (16) is called pseudo-likelihood. Unfortunately, the pseudo-likelihood function provides inconsistent parameter estimates if $\Pr(y_i = |\mathbf{x}_i)$ is misspecified. The selection of the assumption on the modeling structure becomes more important to obtain consistent estimates using pseudo-likelihood function than the conventional MLE. As mentioned in the preceding section, the multinomial logit model provides consistent estimates except constant terms when the conventional MLE is applied. However, it is true only when the multinomial logit model well represents the underlying structure. As the model of injury severity, the ordinality of the severity is a natural structure. The lack of the ordinality in multinomial logit model theoretically creates a higher risk of obtaining inconsistent parameter estimates when the pseudo-likelihood function is applied compared with ordered-response probit model and sequential binary probit model. However, it should be noted that ordinality is in some cases a perceived effect. It is more likely to have ordinality in severity causing factors such as speed, but not have that ordinality translate into ordinal severity outcomes. For example, higher speeds do not necessarily cause higher severities. In addition, it should be noted that ordinality is a restriction on parameter estimates. It is empirically very clear from the literature on multinomial and nested logit models that parameter restrictions can be severe and unjustified if ordinality is assumed in the specification. So specification search is a tradeoff.

From Eq. (16), the consistency of the coefficient estimates except constant terms for the multinomial logit models can be drawn even when the conventional MLE is applied. The probability of choosing alternative j for the multinomial logit model is given as

$$\Pr(y_i = j_i|\mathbf{x}_i) = \frac{\exp(\beta_{j_i}\mathbf{x}_i)}{\sum_{k=1}^{J}\exp(\beta_k\mathbf{x}_i)}. \tag{17}$$

Applying Eq. (17) into Eq. (16) gives

$$\ln L = \sum_{i=1}^{N} \ln \frac{[H(j_i)/Q(j_i)](\exp(\beta_{j_i}\mathbf{x}_i)/\sum_{k=1}^{J}\exp(\beta_k\mathbf{x}_i)}{\sum_{k=1}^{J}\left\{[H(k)/Q(k)](\exp(\beta_k\mathbf{x}_i)/\sum_{k=1}^{J}\exp(\beta_k\mathbf{x}_i)\right\}}$$

$$= \sum_{i=1}^{N} \ln \frac{[H(j_i)/Q(j_i)]\exp(\beta_{j_i}\mathbf{x}_i)}{\sum_{k=1}^{J}\{[H(k)/Q(k)]\exp(\beta_k\mathbf{x}_i)\}}$$

$$= \sum_{i=1}^{N} \ln \frac{\exp(\beta_{j_i}\mathbf{x}_i + \alpha_{j_i})}{\sum_{k=1}^{J}\exp(\beta_k\mathbf{x}_i + \alpha_k)}, \tag{18}$$

where $\alpha_k = \ln[H(k)/Q(k)]$. Eq. (18) suggests that the conventional MLE provides consistent coefficient estimates except constant terms for the multinomial logit models, and the constant terms are biased by α_k. The true constant terms and α_k cannot be separated, so the true constant and $Q(j)$ cannot be estimated separately. This is another disadvantage of the multinomial logit models when the pseudo-likelihood function is applied.

Assuming accidents of the highest severity J have no underreporting, the reporting rate of the accidents of severity j can be calculated by

$$\frac{H(j)/Q(j)}{H(J)/Q(J)}. \tag{19}$$

In the case where severities other than the highest level are assumed to have no underreporting, the ratio of the unknown population shares of those severities (in Eq. (19)) without underreporting are fixed at the ratio of the sample shares.

Applying Eq. (3) into Eq. (16), the estimator provides the estimates of $\boldsymbol{\beta}$, μ_js and $Q(j)$s for the ordered-response probit model. On the other hand, for the sequential binary probit model, separate estimation can be conducted using the pseudo-likelihood function. For separate estimations of the sequential binary probit model, $H(j)$'s and $Q(j)$'s are defined as the sample and population shares, respectively among the severities included in the likelihood function.[3]

In this study, the ordered-response probit model and the sequential binary probit model are estimated, both of which assume that the error terms follow normal distributions. The coefficient estimates of the two models cannot be compared directly because the structures of the models are different. In order to compare the estimated impact of explanatory variables on the injury severity between the models, their elasticities are calculated. The elasticity is used to measure the effect that a 1% change in explanatory variable will have on injury severity probability. Thus, it is valid for continuous variables only, and not applicable for indicator variables that take on values of zero or one only. The average pseudo-elasticity is used for binary indicator variables in this study, which gives the average rate of change in probability when a variable is changed from zero to one. The average pseudo-elasticity of the indicator variable x_k for severity j is given as

$$E_{x_k}^{P(j)} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \frac{\Pr(j|x_{i1},\ldots,x_{i,k-1},1,x_{i,k+1},\ldots,x_{iK}) - \Pr(j|x_{i1},\ldots,x_{i,k-1},0,x_{i,k+1},\ldots,x_{iK})}{\Pr(j|x_{i1},\ldots,x_{i,k-1},0,x_{i,k+1},\ldots,x_{iK})}, \tag{20}$$

where K is the number of explanatory variables, and $\tilde{N}$ is the sample of accidents with $x_{\cdot,k} = 0$. The pseudo-elasticity concept is appropriate for injury severity models that are conditioned on an accident having happened (Shankar and Mannering, 1996; Chang and Mannering, 1999). In such structures, the majority of variables correlated with injury severity are indicator variables (with values of "0" or "1" for the presence or absence of a condition associated with the variable in question).

4. Data

The data set used for the empirical analysis in this study is obtained from the Washington State Highway accident records database, which is supplied by the Washington State Department of Transportation. The database contains information on single and multi-vehicle accident types. For this study, data on single-vehicle accidents involving fixed objects were extracted from the database for the period occurring January 1993 to July 1996. The data set is a bit dated, so newer data sets are desired for the empirical investigation of the effects of contributing factors on injury severity. However, the data set used in the study is well prepared and by no means insufficient for the investigation of the underreporting bias. Also, the results of this study provide a good baseline when the studies with newer data sets are carried out. In the United States, collisions with fixed objects account for 16% of all reported accidents, and they result in 31% of all fatal accidents (National Highway Traffic Safety Administration, 2003); hence the data set includes a considerably larger number of severe accidents than datasets with other accident types. In addition, single-vehicle accidents are more likely to be underreported than multi-vehicles accidents.

The records contain a variety of information including weather conditions, roadway conditions, object type, driver's conditions, vehicle conditions as well as injury severity reported at the time of the accident. The explanatory variables used in the empirical analysis of this study are summarized in Table 1. The severity of the most severely injured person and that of the driver are both recorded. In this study, injury severity of the driver is analyzed. An analysis of most severe injury is a logical extension, but with additional constraints relating to the partial observability in information relating to most severely injured occupant: The most severe injury severity in the accident and the number of persons injured in the accident and that of fatality are recorded, so the information on passenger injury severity for every passenger is not consistently available (Yamamoto and Shankar, 2004). Injury severity is numerically categorized in this study as (1) property damage only, (2) possible injury, (3) evident injury, (4) disabling injury, and (5) fatality. The sample distributions of driver injury severity by urban and rural location type are summarized in Table 2. The Yamamoto-Shankar study (2004) provides statistical justification for separating injury analyses by location type. Generally speaking, parameters are not transferable between urban and rural locations on the basis of likelihood ratio tests. In both urban and rural area, property damage only accidents are dominant, while evident injury accidents contribute a greater percentage than possible injury accidents in rural locations. It is noted that the number of possible injury accidents is almost the same as that of evident injury accidents in urban area, and smaller in rural area than that of evident injury accidents. One possible reason for this is larger underreporting bias of possible injury accidents in rural area than that in urban area as literature suggested (e.g., Amoros et al., 2006).

5. Results

The models were estimated by using a matrix programming language, GAUSS, where the pseudo-likelihood function was coded by the authors and the routine for maximum likelihood estimation provided by GAUSS was used. Results from the underreporting

[3] A severe computational difficulty with the estimator is pointed out by Imbens (1992) in cases where large population shares over multiple levels of severity are involved.

Table 1
Description of explanatory variable

Variable	Description	Mean
Road condition		
Intersection indicator	1 if accident occurred at intersection and related, 0 otherwise	0.0693
Intersection-related indicator	1 if accident occurred at intersection-related but not at intersection, 0 otherwise	0.0068
Icy roadway surface indicator	1 if roadway surface condition was ice or snow, 0 otherwise	0.2355
Off roadway indicator	1 if accident occurred off roadway, 0 otherwise	0.9656
Rain indicator	1 if the weather was raining, 0 otherwise	0.2068
Fixed-object type		
Post indicator	1 if accident occurred in collision with wood and metal sign post, guide post, 0 otherwise	0.0427
Ditch indicator	1 if accident occurred in collision with culvert end or other appurtenance in ditch, roadway ditch, 0 otherwise	0.0797
Guardrail end indicator	1 if accident occurred in collision with leading end of guardrail, 0 otherwise	0.0113
Guardrail face indicator	1 if accident occurred in collision with face of guardrail, 0 otherwise	0.1564
Concrete barrier face indicator	1 if accident occurred in collision with face of concrete barrier, 0 otherwise	0.1827
Bridge face indicator	1 if accident occurred in collision with face of bridge, 0 otherwise	0.0441
Tree indicator	1 if accident occurred in collision with tree or stump, 0 otherwise	0.0529
Fence indicator	1 if accident occurred in collision with fence, 0 otherwise	0.0356
Restraint system		
Lap belt indicator	1 if lap belt was used, 0 otherwise	0.0694
Shoulder belt indicator	1 if shoulder belt was used, 0 otherwise	0.0079
Lap and shoulder belt indicator	1 if lap and shoulder belt are used, 0 otherwise	0.7798
Air bag and belt indicator	1 if air bag and belt are used, 0 otherwise	0.0147
Vehicle and driver		
Vehicle age	In 10 years	0.9219
Large truck indicator	1 if vehicle was truck over 10K pounds, 0 otherwise	0.0026
Truck indicator	1 if vehicle was truck tractor, semi-trailer or other truck combinations, 0 otherwise	0.0331
Motorcycle indicator	1 if vehicle is motorcycle, scooter bike or moped, 0 otherwise	0.0044
Defective brake indicator	1 if brakes were defective, 0 otherwise	0.0110
Tire blow indicator	1 if tires had punctured or blown, 0 otherwise	0.0140
Male driver indicator	1 if driver was male, 0 otherwise	0.6608
Driver's age	In 100 years	0.3427
Over speed limit indicator	1 if speed exceeded stated speed limit, 0 otherwise	0.0277
Exceed safety speed indicator	1 if speed exceeded reasonable safe speed, 0 otherwise	0.4364
Improper turning indicator	1 if driver made improper turning, 0 otherwise	0.0075
Asleep indicator	1 if driver was apparently asleep, 0 otherwise	0.0908
Driver sobriety indicator	1 if driver had been drinking and ability was impaired, 0 otherwise	0.1518
Number of passengers		0.5076

methodologies described above for ordered probit and sequential binary probit are presented in this paper. We present results for the case in which underreporting without known population shares is assumed to occur in property damage only and possible injury accidents. Since these two severities occupy the lower spectrum of the accident severity scale and comprise over 75% of reported accidents, we confine our attention to parameter biases induced by underreporting in these severity types. Other cases involving underreporting possibilities in higher severities were also explored in this study but not reported for the purpose of brevity. We note that in applying the pseudo-likelihood function to account for underreporting, we exclude the impact of error correlations to simplify the computational procedure.[4]

We present summary statistics of overall model fit as shown in Table 3. The results show that both for urban and rural area accidents, the sequential binary probit model has a better log-likelihood values at convergence and the log-likelihood ratio statistics, suggesting that it is an equally if not more plausible model. For the purpose of brevity, the coefficient estimates of the ordered-response probit model and the sequential binary probit model for urban area accidents are presented in Table 4.

Table 2
Sample distribution of driver injury severity by urban and rural accident types

	Urban	(%)	Rural	(%)
Property damage only	6125	(63)	6514	(61)
Possible injury	1646	(17)	1357	(13)
Evident injury	1602	(16)	2191	(21)
Disabling injury	297	(3)	474	(4)
Fatality	53	(1)	104	(1)
Total	9723	(100)	10,640	(100)

The explanatory variables are selected based on their t-statistics, and those with 95% or higher levels of significance are included in the final models. In the sequential binary probit model, a majority of the factors significantly correlate with β_1, the vector associated with property damage only or higher severity levels. The number of factors significantly correlated with higher severity levels typically declines as evidenced in the size of the vectors β_2, β_3 and β_4. It implies that the factors affecting the injury severity change according to the level of the severity itself. This assumption was described in Section 3, and identified as a methodological flexibility in the sense that it does not constrain parameters to be the same across all levels of severity.

[4] We remind the reader here that the ordered-response model was proposed (McCarthy and Madanat, 1994) in order to account for error correlations. Our analysis of error correlations suggests that given the sample size, notable correlation is present only among property damage only and possible injury crashes. Statistical significance of the correlation using a t-test is −2.40 for urban area accidents and −4.37 for rural area accidents. The results appear consistent with nested logit structures that suggest shared unobservables between property damage only and possible injury accident types (see for example Shankar et al., 1996; Holdridge et al., 2005). However, we note here the weaker significance of the urban area correlation measure.

Table 3
Summary statistics of the model estimation using pseudo-likelihood function accounting for underreporting

	Urban (N = 9723)				Rural (N = 10,640)			
	LL(C)	LL(β)	ρ^2	$\bar{\rho}^2$	LL(C)	LL(β)	ρ^2	$\bar{\rho}^2$
Ordered-response probit model	−9955	−9193	0.077	0.073	−11,409	−10,288	0.098	0.095
Sequential binary probit model	−9955	−9094	0.086	0.082	−11,409	−10,212	0.105	0.099

LL(C) and LL(β) stand for log-likelihood computed with only a constant term and at convergence, respectively. ρ^2 and $\bar{\rho}^2$ stand for log-likelihood ratio statistic and the adjusted log-likelihood ratio statistic for degree of freedom, respectively. The higher statistics represent better model for both statistics.

5.1. *Coefficient estimates*

5.1.1. *Road condition variables*

Intersection indicator, intersection-related indicator, icy roadway surface indicator and rain indicator have negative coefficient estimates in the ordered-response probit model, suggesting higher probability of lower injury severity in fixed-object accidents. Similar results are found in the sequential binary probit model, but the effects are not statistically significant on the higher severity levels. Off-the-roadway indicator has a positive coefficient estimate and the largest elasticity in the ordered-response probit model. On the other hand, the off roadway indicator has the same positive coefficient estimate as the factor determining whether the severity level is property damage only or higher, but a smaller elasticity for fatality in the sequential binary probit model. The large elasticity in the ordered-response probit model might be caused by the unrealistic assumption of the ordered-response probit model that the effects of the factors are constrained to be the same regardless of the severity levels.

5.1.2. *Fixed-object type variables*

Type of fixed objects is associated with injury severity significantly as well. Collisions with leading ends of guardrail and trees are associated with higher injury severity in the ordered-response probit model, while collisions with sign posts, appurtenances in ditch, faces of guardrail, concrete barrier or bridge, and fences is associated with less severe injury. The same directional effect is found in the sequential binary probit model, but the levels of the severity that the factors are associated are estimated as varying across the type of fixed objects. As an example, guardrail end indicator has significant positive coefficient estimates as the factor determining whether the severity level is property damage only or higher and as the factor determining whether the severity level is disabling injury or fatality. The estimated elasticity of the guardrail end indicator for fatality is much higher than that in the ordered-response probit model.

5.1.3. *Restraint system variables*

On the effects of safety restraint system use, the results suggest that proper use of the restraint system tends to be associated with less severe injury. Especially, the lap-and-shoulder-belt use indicator has the highest t-static values among the explanatory variables in the ordered-response probit model. Similar results are found in the sequential binary probit model in association with severity levels involving possible injury or higher as well as severity levels involving evident injury or higher. The results imply that the proper use of lap and shoulder belt decreases the injury severity regardless of the severity levels. The elasticity of the lap and shoulder belt for fatality is estimated as about −90% in both models.

5.1.4. *Vehicle and driver variables*

Several vehicle condition variables including vehicle age, type of the vehicle, defective brake and tire blow indicators are estimated as associated with injury severity in the ordered-response probit model as well as the sequential binary probit model. However, the effects are limited in the binary probit model to factors determining whether the severity level is property damage only or higher. This might imply that there exist many cases where defective brake and tire blow are the primary reason for the accidents especially for the least severe accident, property damage only.

Sex of driver (female), older driver, exceeding speed limit, exceed safe speed for driving conditions, asleep at the time of the accident, and sobriety indicators are estimated to be associated strongly with higher injury severities, while improper turning and the number of passengers variables are associated with lower injury severities. Of these, driver sobriety has the largest elasticity in both models, and significant positive coefficients. The results imply that driver under the influence of alcohol increases the injury severity significantly in almost all levels of the injury severity. The negative effect of the number of passenger on injury severity may imply that drivers tend to drive carefully when they are driving with more passengers, but the causal relationship should be further investigated in future study.

5.2. *Underreporting*

The estimated reporting rates for property damage only and possible injury by ordered-response probit model and sequential binary probit model are shown in Table 5. The point estimates and confidence intervals are shown in the table. The former is calculated by applying estimated $Q(j)$ into Eq. (19), and the latter is calculated from the standard error of the estimated reporting rate which is calculated by using the estimated standard error of $Q(j)$ and applying the delta method (Greene, 2003). The results show that the point estimates of the reporting rates are well below 100% for both property damage only and possible injury regardless of the modeling structures. Also, the results suggest that the reporting rate is lower for possible injury than for property damage only in both models. This may sound a bit counter intuitive, but is consistent with the literature. As shown in Table 6, Elvik and Mysen (1999) suggest that very slight injury crashes have a lower reporting rate than property damage only crashes. They used a different term, very slight injury, from ours, but the possible injury outcome as defined in our study is included in this category. Elvik and Mysen provide average, lowest and highest reporting rates for each injury severity level, and our point estimates by sequential binary probit model fall between lowest and highest cases, while those by ordered-response probit model have higher reporting rates than the highest case. However, the 95% confidence interval of our estimates is wide except for possible injury by sequential binary probit model. Thus we cannot judge the size of the underreporting for each injury severity from the results. At least, the results statistically significantly suggest that there exists underreporting for possible injury accidents regardless of the assumed modeling structures. For rural accidents, the estimates of the reporting rates turned out to be rather vague than those for urban accidents, though the results are omitted for brevity, thus we could not obtain any meaningful insights on the difference in underreporting between urban and rural accidents.

Table 4
Coefficient estimates for urban models using pseudo-likelihood function assuming accidents with property damage only and possible injury include underreporting

	Ordered-response probit model			Sequential binary probit model								
	$\boldsymbol{\beta}$		E[a] (%)	$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\beta}_3$		$\boldsymbol{\beta}_4$		E[a] (%)
	Coefficient	*t*-Statistics		Coefficient	*t*-Statistics	Coefficient	*t*-Statistics	Coefficient	*t*-Statistics	Coefficient	*t*-Statistics	
Constant	0.485	1.59		1.203	1.25	−1.667[b]	−17.06	−0.745	−11.8	−1.475	−6.2	
Road condition												
Intersection indicator	−0.113	−2.8	−31	−0.147	−2.82							−7
Intersection-related indicator	−0.347	−2.93	−69	−0.410	−2.37							−21
Icy roadway surface indicator	−0.166	−4.67	−42	−0.182	−3.33	−0.107	−2.29					−29
Off roadway indicator	0.518	5.94	524	0.677	5.64							57
Rain indicator	−0.117	−4.5	−32	−0.123	−2.78	−0.108	−3.03					−27
Fixed-object type												
Post indicator	−0.375	−5.68	−72	−0.463	−4.44							−24
Ditch indicator	−0.244	−4.51	−56	−0.211	−2.56	−0.309	−4.05					−58
Guardrail end indicator	0.204	2.23	92	0.244	1.86					1.257	2.56	535
Guardrail face indicator	−0.149	−4.38	−39	−0.169	−3.16	−0.101	−2.29					−27
Concrete barrier face indicator	−0.112	−4.09	−31	−0.109	−2.80	−0.132	−3.82			0.411	2.07	45
Bridge face indicator	−0.11	−2.21	−31	−0.135	−2.01							−7
Tree indicator	0.113	1.83	44							1.023	3.46	352
Fence indicator	−0.24	−3.11	−56	−0.326	−3.02							−16
Restraint system												
Lap belt indicator	−0.647	−8.46	−90	−0.778	−3.77			−0.666	−3.83			−82
Shoulder belt indicator	−0.489	−3.79	−82	−0.689	−3.20							−36
Lap and shoulder belt indicator	−0.836[b]	−12.3	−92	−0.963	−4.31	−0.374[b]	−9.82	−0.373	−5.23	−0.578	−3.14	−93
Air bag and belt indicator	−0.372	−4.58	−72	−0.394	−2.90							−20
Vehicle and driver												
Vehicle age	−0.006	−3.29	−17	−0.007	−3.02							−3
Large truck indicator	−0.603	−2.17	−88	−0.730	−2.14							−38
Truck indicator	−0.159	−1.74	−41	−0.275	−2.51							−14
Motorcycle indicator	−0.502	−2.09	−83	−0.683	−1.99							−36
Defective brake indicator	−0.225	−2.10	−53	−0.272	−1.99							−14
Tire blow indicator	−0.205	−2.11	−50	−0.297	−2.50							−15
Male driver indicator	−0.215	−7.6	−50	−0.334	−5.19	0.123	3.97					14
Driver's age	0.001	1.8	15							0.012	2.23	72
Over speed limit indicator	0.161	2.47	68									
Exceed safety speed indicator	0.075	2.78	28	0.072	2.01							4
Improper turning indicator	−0.376	−2.74	−72	−0.483	−2.26							−25
Asleep indicator	0.242	5.02	117	0.228	3.06	0.242	3.67					89
Driver sobriety indicator	0.367	9.21	222	0.319	4.18	0.406	8.98	0.18	2.50			268
Number of passengers	−0.041	−3.09	−13	−0.067	−3.38							−3
μ_2	1.338[b]	4.94										
μ_3	2.261[b]	9.49										
μ_4	3.039[b]	12.58										
Sample size		9723			9723		3598		1952		350	
LL(0)	−15648.6			−6739.5		−2493.9		−1353.0		−242.6		
LL(β)	−9193.3			−5799.6		−2269.2		−892.6		−132.7		

[a] Pseudo-elasticity is shown for fatality except vehicle age and driver's age for which conventional elasticity.

[b] Coefficient estimate is significantly different at 95% level of significance from that estimated by conventional MLE.

Table 5
Estimation results of reporting rate by injury severity for our data

	Ordered-response probit model		Sequential binary probit model	
	Point estimate (%)	95% confidence interval (%)	Point estimate (%)	95% confidence interval (%)
Property damage only	73.4	−37.7 to 184.4	27.5	−78.5 to 133.5
Possible injury	24.0	−2.4 to 50.3	2.6	1.6 to 3.6

5.3. Underreporting bias

Parameter bias from underreporting in property damage only and possible injury accidents is shown in Table 7. We refer to the model without underreporting as the “naïve” model. This reference is used to illustrate the extent of parameter bias that is induced when underreporting is not taken into account. For the purpose of brevity, the results shown are for urban area accidents. As Table 7 shows significant parameter bias occurs when underreporting is not taken into account. Since property damage only and possible injury accidents are the cases where underreporting is assumed to occur, we find that in the ordered-response model, mean non-constant bias is approximately 23.4%. Biases vary from as low as 11% for the defective brake indicator to as high as 100% for driver age. Mean elasticity bias is approximately 14% with a high of 38% for off-roadway accident occurrence indicator to 3% for several factors.

Parameter bias is significantly varied in the sequential binary probit model since the parameter vectors change according to the severity thresholds. For the property damage only or higher parameter vector, mean parameter bias is approximately 4.49%, while for possible injury or higher parameters, the bias increases to 43.1%. Mean elasticity bias is approximately 45.5% in the sequential binary probit model. The higher parameter bias for possible injury results from estimated low reporting rate for possible injury severity accidents. On the whole, the results suggest that the higher underreporting bias causes higher parameter and elasticity biases.

Table 6
Results of meta-analysis of reporting rate by injury severity (Elvik and Mysen, 1999)

	Average (%)	Lowest case (%)	Highest case (%)
Property damage only	25	22	38
Very slight injury	11	1	17
Slightly injury	27	6	76
Seriously injured	69	47	97
Fatality	95	87	106

6. Summary and conclusion

Many studies have applied ordered-response probit models and sequential binary logit models to examine injury severity in rela-

Table 7
Parameter bias in naïve model compared to unbiased model accounting for underreporting in property damage and possible injury accidents

	Ordered probit		Sequential binary probit		
	Parameter bias[a]	Elasticity bias[a]	Property damage or higher parameter bias[a]	Possible injury or higher parameter bias[a]	Elasticity bias[a]
Constant	−56		−70	−122	
Intersection indicator	21	16	1		114
Intersection-related indicator	18	9	5		90
Icy roadway surface indicator	17	12	3	59	0
Off roadway indicator	16	38	−3		140
Rain indicator	20	13	8	45	−11
Post indicator	19	7	2		79
Ditch indicator	18	9	8	46	−16
Guardrail end indicator	23	27	14		22
Guardrail face indicator	21	13	7	35	0
Concrete barrier face indicator	18	10	6	41	27
Bridge face indicator	29	19	6		100
Tree indicator	32	34			0
Fence indicator	24	11	3		100
Lap belt indicator	25	4	14		10
Shoulder belt indicator	18	5	13		78
Lap and shoulder belt indicator	22	3	12	37	0
Air bag and belt indicator	21	8	14		105
Vehicle age	17	18	0		133
Large truck indicator	18	3	−2		58
Truck indicator	20	12	−10		71
Motorcycle indicator	27	7	5		67
Defective brake indicator	11	6	0		93
Tire blow indicator	18	8	−8		80
Male driver indicator	24	12	3	46	−229
Driver’s age	100	27			0
Over speed limit indicator	23	25			
Exceed safety speed indicator	20	18	6		100
Improper turning indicator	30	11	6		84
Asleep indicator	19	22	4	46	−28
Driver sobriety indicator	20	29	7	33	−37
Number of passengers	20	8	3		133
μ_2	−57				
μ_2	−26				
μ_4	−16				

[a] Percent bias in parameter computed as the percent change in parameter value with respect to the unbiased parameter.

tion to driver factors, vehicle attributes, and roadway conditions. These studies have implicitly assumed random sampling when using conventional MLE. The literature, however, suggests that lower severity accidents have lower reporting rates, which results in the outcome-based sampling of traffic accident data. This study attempts to investigate the effects of the underreporting on the parameter estimates. This study also compares parametric insights on key roadway, environmental, vehicle and driver related factors through the application of ordered response and sequential binary probit models when accounting for underreporting.

Underreporting results suggest that the estimates of the effect of the explanatory variables can be significantly biased if underreporting is not considered in the parameter estimation. Parameter elasticity biases are also found to be significant. While the ordered-probit model shows a wider variation in biases, biases are found to be significantly higher in the sequential binary probit model for severity levels involving possible injury or higher. This suggests interesting conclusions. While underreporting is expected as the largest in property damage only accidents, the mean parameter bias appears to be minimal, only around 4.5%; whereas when underreporting is considered for possible injury accidents, the mean bias significantly increases. However, in light of the fact that nearly 33% of all factors are found to be strongly associated with higher severities, parameter bias becomes a significant issue if underreporting is not considered. Environmental effects, safety restraint use, driver sex, driver fault and fixed object type factors are all significantly biased.

This study is the first step in the statistical investigation of the size and elasticity of factors affecting injury severity under underreporting. Further research is needed to confirm the efficiency of the estimator for large-scale accident analysis.

Acknowledgements

The conception of this study was brought by the discussion during the first author's stay at the University of Washington, Seattle as a visiting scholar hosted by Fred L. Mannering in 2000–2001. This study was in part supported through a research grant by Nagoya University Foundation for Promotion of Science.

References

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accid. Anal. Prev. 37, 417–425.

Amoros, E., Martin, J.-L., Laumon, B., 2006. Under-reporting of road crash casualties in France. Accid. Anal. Prev. 38, 627–635.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on accident frequencies and severities. Accid. Anal. Prev. 33, 99–109.

Chang, L.-Y., Mannering, F., 1999. Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accident. Accid. Anal. Prev. 31, 579–592.

Cosslett, S.R., 1981a. Efficient estimation of discrete-choice methods. In: Manski, C., McFadden, D. (Eds.), Structural Analysis of Discrete Choice Data with Econometric Applications. MIT Press, Cambridge, MA, pp. 51–111.

Cosslett, S.R., 1981b. MLE for choice-based samples. Econometrica 49, 1289–1316.

Cosslett, S.R., 1993. Estimation from endogenously stratified samples. In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics, vol. 11. Elsevier, North-Holland, pp. 1–43.

Cosslett, S.R., 2005. Efficient semiparametric estimation from a smoothed likelihood function. In: Econometric Society World Congress, London, August.

Dissanayake, S., Lu, J.J., 2002a. Factors influential in making an injury severity difference to older drivers involved in fixed object-passenger car crashed. Accid. Anal. Prev. 34, 609–618.

Dissanayake, S., Lu, J., 2002b. Analysis of severity of young driver crashes: sequential binary logistic regression modeling. Transport. Res. Rec. 1784, 108–114.

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

Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. Transport. Res. Rec. 1665, 133–140.

Greene, W.H., 2003. Econometric Analysis, fifth edition. Prentice Hall, Upper Saddle River, NJ.

Hauer, E., Hakkert, A.S., 1988. Extent and some implications of incomplete accident reporting. Transport. Res. Rec. 1185, 1–10.

Holdridge, J.M., Shanker, V.N., Ulfarsson, G.F., 2005. The crash severity impacts of fixed roadside objects. J. Safety Res. 36, 139–147.

Imbens, G.W., 1992. An efficient method of moments estimator for discrete choice models with choice-based sampling. Econometrica 60, 1187–1214.

Khattak, A.J., 2001. Injury severity in multi-vehicle rear-end crashes. Transport. Res. Rec. 1746, 59–68.

Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34, 313–321.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accid. Anal. Prev. 34, 149–161.

McCarthy, P.S., Madanat, S., 1994. Highway accident data analysis: alternative econometric methods. Transport. Res. Rec. 1467, 44–49.

National Highway Traffic Safety Administration, 2003. Traffic Safety Facts 2003: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, DC.

O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accid. Anal. Prev. 28, 739–753.

Saccomanno, F.F., Nassar, S.A., Shortreed, J.H., 1996. Reliability of statistical road accident injury severity models. Transport. Res. Rec. 1542, 14–23.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. J. Safety Res. 27 (3), 183–194.

Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accid. Anal. Prev. 28, 391–401.

Shankar, V., Albin, R., Milton, J., Nebergall, M., 2000. In-service performance-based roadside design policy: preliminary insights from Washington State's bridge rail study. Transport. Res. Rec. 1720, 72–79.

Shimamura, M., Yamazaki, M., Fujita, G., 2005. Method to evaluate the effect of safety belt use by rear seat passengers on the injury severity of front seat occupants. Accid. Anal. Prev. 37, 5–17.

Ulfarsson, G.F., Mannering, F.L., 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accid. Anal. Prev. 36, 135–147.

Yamamoto, T., Shankar, V.N., 2004. Bivariate ordered-response probit model of driver's and passenger's injury severities in collision with fixed objects. Accid. Anal. Prev. 36, 869–876.