Accident Analysis and Prevention 95 (2016) 57–66

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# A method to account for and estimate underreporting in crash frequency research

Jonathan S. Wood [a,*], Eric T. Donnell [b], Christopher J. Fariss [c]

[a] Department of Civil and Environmental Engineering, South Dakota State University, Crothers Engineering Hall, Box 2219, Brookings, SD 57007, United States
[b] Department of Civil and Environmental Engineering, The Pennsylvania State University, 231 Sackett Building, University Park, PA 16802, United States
[c] Department of Political Science, The Pennsylvania State University, 227 Pond Lab, University Park, PA 16802, United States

ARTICLE INFO

*Article history:*
Received 17 November 2015
Received in revised form 9 June 2016
Accepted 17 June 2016

*Keywords:*
Crash underreporting
Negative binomial underreporting
Poisson underreporting with heterogeneity
Random parameters negative binomial
Crash frequency
Predictive modeling

ABSTRACT

Underreporting is a well-known issue in crash frequency research. However, statistical methods that can account for underreporting have received little attention in the published literature. This paper compares results from underreporting models to models that account for unobserved heterogeneity. The difference in the elasticities between the negative binomial underreporting model and random parameters negative binomial models, which accounts for unobserved heterogeneity in crash frequency models, are used as the basis for comparison. The paper also includes a comparison of the predicted number of unreported PDO crashes based on the negative binomial underreporting model with crashes that were reported to police but were not considered reportable to PennDOT to assess the ability of the underreporting models to predict non-reportable crashes.

The data used in this study included 21,340 segments of two-lane rural highways that are owned and maintained by PennDOT. Reported accident frequencies over an eight year period (2005–2012) were included in the sample, producing a total of 170,468 segment-years of data. The results indicate that if a variable impacts both the true accident frequency and the probability of accidents being reported, statistical modeling methods that ignore underreporting produce biased regression coefficients. The magnitude of the bias in the present study (based on elasticities) ranged from 0.00–16.79%. If the variable affects the true accident frequency, but not the probability of accidents being reported, the results from the negative binomial underreporting models are consistent with analysis methods that do not account for underreporting.

Published by Elsevier Ltd.

## 1. Introduction

Underreporting of undesirable events, such as accidents (e.g., industrial accidents, worker-related accidents, traffic accidents, etc.), is a well-documented issue (Brookoff et al., 1993; Kamura and Chin, 2005; Kemp, 1973; Leigh et al., 2004; Lord and Mannering, 2010; Probst and Estrada, 2010; Probst and Graso, 2013; Probst et al., 2013). A growing body of literature has considered the impacts of underreporting crash severity models (Patil et al., 2012; Yamamoto et al., 2008; Yasmin and Eluru, 2013; Ye and Lord, 2006). However, consideration of the impacts of underreporting on statistical inference in crash frequency analysis has received little attention (Hauer and Hakkart, 1988; Hauer, 2006; Kamura and Chin, 2005; Kemp, 1973; Ma and Li, 2010). Traffic crash reporting depends on several factors, including:

1. The level of vehicle damage, which is often used as a measure to determine if a crash event is reportable (Hauer, 2006),
2. the most severe level of injury among the driver(s) or passengers involved in the crash (i.e., more severe crashes, such as fatal or severe injuries, are more likely to be reported) (Kemp, 1973; Patil et al., 2012; Yamamoto et al., 2008; Yasmin and Eluru, 2013),
3. the willingness of those involved in the crash to report the crash to the police, which may be influenced by insurance cost considerations (Hosios and Peters, 1989),
4. the willingness of the responding officer to file a crash report (e.g., if the officer judges the level of damage to be significant enough), and
5. the accuracy of reporting the crash with regards to the location, severity, and other factors.

* Corresponding author.
*E-mail addresses:* jsw277@psu.edu (J.S. Wood), edonnell@engr.psu.edu (E.T. Donnell), cjf20@psu.edu (C.J. Fariss).

http://dx.doi.org/10.1016/j.aap.2016.06.013
0001-4575/Published by Elsevier Ltd.

These factors clearly indicate that crash counts are underreported due to multiple non-random factors leading to selection bias in the reported crashes.

In a summary of research attempting to estimate the levels of crash underreporting, estimates found ranged from 11 to 65% for all crashes, 46–62% for non-injury (property damage only) crashes, 7–80% for injury crashes, and 0–9% for fatal crashes (Hauer and Hakkart, 1988). Different underreporting rates among the various severity levels is intuitive because legal and financial issues lead to many non-injury and minor injury crashes not being reported. For example, these low severity crashes may go unreported because there is not enough vehicle damage sustained in the crash, or drivers may fear increased insurance costs if the crash is reported. For fatal crashes, underreporting is unlikely but may happen if there are errors in reporting the crash location, or if there is a lack of follow-up to know whether a fatality has occurred after an individual involved in a crash has left the crash location (crashes are considered fatal if anyone involved in the crash dies within 30 days of the crash due to crash-related injuries (National Highway Traffic Safety Administration, 2014)).

More recent research has attempted to estimate the levels of underreporting based on the injury severity level by combining crash and hospital data (Abay, 2015; Alsop and Langley, 2001; Amoros et al., 2006; Elvik and Mysen, 1999; Rosman and Knuiman, 1994). These studies have provided evidence of underreporting, with levels similar to those reported by Hauer and Hakkart (1988), but have not provided solutions for dealing with underreporting in crash frequency modeling.

Due to the correlation between the severity of the crash and the probability of it being reported, the determination of whether changes in the number of reported crashes is due to changes in the crash severity distribution or changes in the actual number of crashes that occurred (or a mixture of both) is unaccounted for in the majority of crash frequency research (i.e., limited to modeling crash frequency by severity if accounted for at all). Underreporting has been accounted for in multiple crash severity studies (Kockelman and Kweon, 2002; Patil et al., 2012; Quddus et al., 2010; Yamamoto et al., 2008; Ye and Lord, 2006). However, only three research articles were found that attempt to model underreporting in crash frequency models (Kamura and Chin, 2005; Ma and Kockelman, 2006; Ma and Li, 2010). One of these studies used a maximum likelihood approach (Kamura and Chin, 2005) while the other two studies used Bayesian estimation methods to estimate Poisson underreporting models (Ma and Kockelman, 2006; Ma and Li, 2010).

The purpose of this paper is to account for underreporting in the development of crash frequency prediction models using two-lane rural highway data from Pennsylvania. The results are compared to commonly used models of crash frequency. This is done by using both Poisson underreporting models with random intercepts and negative binomial underreporting models. The underreporting model results are then compared to the most common type of regression in traffic safety that accounts for multiple sources of unobserved heterogeneity (i.e., random parameters negative binomial models that include the same predictor variables) (Mannering et al., 2016), without considering underreporting. Finally, the underreporting models for property damage only (PDO) crashes are compared with observed non-reportable crashes to ascertain whether the prediction made from underreporting models can be used to predict the levels of crash underreporting.

## 2. Background: heterogeneity in count models

Count regression models have been applied in many fields of research. Recent trends in transportation safety indicate a strong push toward the use of random parameters count models (Mannering and Bhat, 2014). The random parameters are said to capture unobserved heterogeneity, which is explained as the variable with the random coefficient being correlated with one or more unobserved variables which affect the outcome (Mannering and Bhat, 2014). However, the random parameters may also be picking up other sources of heterogeneity such as incorrect functional form (Mannering et al., 2016), missing important interactions, or measurement error.

One potential source of unobserved heterogeneity that the random parameters model may capture is related to underreporting. If there is both a counting process and a reporting mechanism that results in an observed count outcome, a random parameter may indicate that underreporting is associated with the variable if the variable is correlated with underreporting (i.e., the random parameter may be picking up the incorrect functional relationship between the crash counts and the variable). Thus, if the latent underreporting process was modeled, the variable would be a predictor of the probability that a crash was reported. When a regression model incorporates both a reporting and count model, the latent reporting process is approximated (providing a model that can be used to predict the number of unreported crashes). Although there is no guarantee that these models perform better or are more useful than random parameters models for predicting observed counts, they may be useful to practicing engineers in predicting unreported crash counts. This is an issue of model validation that is investigated in this paper.

Another potential issue related to underreporting of crashes is that when a variable affects both the number of crashes and the probability of crash reporting, the regression estimate for that variable is biased due to endogeneity if the mechanism for crash reporting is not accounted for (since endogeneity occurs whenever one or more predictor variables are correlated with the error term (Greene, 2011; Kennedy, 2008)). Even when the parameter is modeled as a random coefficient, the estimate may not be good for predictive purposes. Given that regression estimates for crash frequency are often used for prediction in transportation engineering, developing models that provide accurate predictions is of great importance.

Regression methods that model the latent reporting process, along with the counting process, have been applied in safety research (Kamura and Chin, 2005; Ma and Li, 2010), but have not been compared with random parameters models or other models that account for unobserved heterogeneity. The majority of safety research that accounts for underreporting of crashes focuses on severity modeling, which does not directly model the latent underreporting process.

## 3. Methodology

Crash frequency prediction models are often developed using negative binomial regression methods to account for overdispersion common in reported crash data (AASHTO, 2010; Lord and Mannering, 2010). These models take the form shown in Eq. (1) (AASHTO, 2010; Wood et al., 2015a).

$$\mu_i = L_i^{\beta_L} \cdot AADT_i^{\beta_{AADT}} \cdot \exp\left(\beta_{\text{int}ercept} + \sum_{j=1}^{J} \beta_j X_j\right) \quad (1)$$

Where $\mu_i$ = the reported crash frequency (i.e., the number of crashes per year for road segment i), $L_i$ = the length of segment i, $\beta_L$ = the estimated coefficient for segment length, $AADT_i$ = the average annual daily traffic for segment i, $\beta_{AADT}$ = the estimated coefficient for AADT, $X_j$ = predictor variable j of J predictor vari-

Download English Version:

https://daneshyari.com/en/article/571912

Download Persian Version:

https://daneshyari.com/article/571912

Daneshyari.com