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

ables, and β_j = the estimated coefficient for variable j of J predictor variables.

Since Eq. (1) is based on reported crash counts, the relationship between the reported and true crash frequency is shown in Eq. (2) (Greene, 2007; Kamura and Chin, 2005; Kemp, 1973; Ma and Kockelman, 2006; Ma and Li, 2010).

$$\mu_i = \mu_{i,true} P_i \tag{2}$$

Where $\mu_{i,true}$ = the true crash frequency for road segment i, and P_i =the average probability that crashes were reported for segment i.

Eq. (2) holds for any level of crash severity being analyzed. Though Eq. (1) is typically used in traffic safety research, such models do not account for underreporting, which could lead to biased regression parameter estimates (Lord and Mannering, 2010). Thus, a method to model the probability of reporting and crash frequency, which captures the reporting process and the true crash frequency, is needed. When accounting for underreporting, Poisson regression is typically used, although negative binomial regression is possible (Cameron and Trivedi, 1998; Winkelmann, 2013). The standard Poisson PDF is shown in Eq. (3), the modified Poisson PDF that accounts for underreporting is shown in Eq. (4), and the modified negative binomial PDF that accounts for underreporting is shown in Eq. (5) (Cameron and Trivedi, 1998).

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{3}$$

$$f(y_i|\mu_{i,true}, P_i) = \frac{e^{-\mu_{i,true}P_i}(\mu_{i,true}P_i)^{y_i}}{y_i!} \tag{4}$$

$$f(y_i|\mu_{i,true}, P_i) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu_{i,true}P_i}\right)^{1/\alpha} \left(1 - \frac{1}{1+\alpha\mu_{i,true}P_i}\right)^{y_i} \tag{5}$$

Where α =the overdispersion parameter.

It should be noted that underreporting models estimate the probability of reporting rather than the probability of underreporting (Cameron and Trivedi, 1998; Winkelmann, 2013). The probability of reporting can be modeled as a binary logit or probit (Cameron and Trivedi, 1998; Winkelmann, 2013). The binary logit is specified as shown in Eq. (6) (Train, 2009).

$$P_i = \frac{\exp(\beta_j Z_j)}{1+\exp(\beta_j Z_j)} \tag{6}$$

Where β_j = a vector of coefficients, and Z_j = a vector of variables.

The Poisson underreporting model can also account for unobserved heterogeneity by allowing the intercept to be random (Greene, 2007). When this is done, the resulting Poisson underreporting model is an overdispersed count model with the function for the observed mean shown in Eq. (7) and variance function for the observed count specified as shown in Eq. (8) (Greene, 2007) The variance function for the negative binomial underreporting model is shown in 9 (Cameron and Trivedi, 1998).

$$\mu_i = \exp\left(\beta_{random} + \sum \beta_j X_j + \varepsilon_i\right)\left(\frac{\exp(\beta_j Z_j)}{1+\exp(\beta_j Z_j)}\right) \tag{7}$$

$$VAR(\mu_i) = \mu_i + (\exp(\sigma^2) - 1)\mu_i^2 \tag{8}$$

$$VAR(\mu_i) = \mu_i + \alpha\mu_i^2 \tag{9}$$

Where β_{random} = the mean value of the random intercept, ε_i =the error term capturing the difference between the mean value of the random intercept and the true intercept for individual i, and σ^2 = the estimated variance of the random intercept.

The underreporting negative binomial regression model cannot be specified with a random intercept as it would have two parameters that account for the unobserved heterogeneity that are perfectly collinear (i.e., σ and α are perfectly collinear due to the Poisson underreporting model with heterogeneity not accounting for the panel nature of the data) (Greene, 2007).

It should be noted that the underreporting Poisson model with heterogeneity, and the negative binomial underreporting models, cannot account for the possibility of correlation between parameters in the model, with the exception of correlation between the error terms in the frequency and reporting portions of the model. This is due to the fact that they do not employ a full random parameters or finite mixture model formulation. It should also be noted that, since the underreporting approach herein estimates a model for a latent process, the results cannot have a causal interpretation. However, if the underreporting model yields predictions for underreporting that are accurate, then the method has potential to have value in transportation safety research and professional practice.

The most common regression method that accounts for unobserved heterogeneity that has been applied to transportation safety is the random parameters negative binomial (Mannering and Bhat, 2014; Mannering et al., 2016). The log likelihood for a random parameters negative binomial model, regardless of the number of random parameters included in the model, is shown in Eq. (10) (Greene, 2007; Anastasopolous and Mannering, 2009).

$$LL = \sum_{\forall i} \ln \int_{\varphi_i} g(\varphi_i) P(n_i|\varphi_i) d\varphi_i \tag{10}$$

Where φ_i= the random distribution for coefficient i, $g(\,)$ = the probability density function of φ_i, and $P(n_i|\varphi_i)$ is the probability for the negative binomial.

The random negative binomial model can account for the panel nature of the data. Thus, the random parameters negative binomial regression model accounts for between-entity unobserved heterogeneity with the random parameters and within-entity unobserved heterogeneity with the overdispersion parameter, while the underreporting Poisson model with heterogeneity combines all sources of unobserved heterogeneity into the random intercept. Underreporting Poisson models are estimated using quadrature-based methods to optimize the maximum likelihood functions while the random parameters models are estimated using halton draw based simulation methods (Greene, 2007). Negative binomial underreporting models can be estimated using maximum likelihood methods (Cameron and Trivedi, 1998). The Poisson underreporting and random parameters models were estimated using the Limdep software. The negative binomial underreporting models were estimated with user-written STATA code (example code is provided in the Supplementary material).

In order to compare the estimates from underreporting and random parameters models, elasticities (for continuous variables) and pseudo-elasticities (for binary indicator variables) were computed. Elasticity is defined as the percent change in the outcome (dependent variable) resulting from a 1% change in the predictor variable. Pseudo-elasticity is the percent change in the outcome when the predictor variable changes from a value of 0 to unity. The elasticities and pseudo-elasticities for count models are estimated using Eqs. (11) and (12), respectively (Greene, 2011; Hilbe, 2011).

$$E = \beta x \tag{11}$$

$$E_{pseudo} = 100(\exp(\beta) - 1) \tag{12}$$

Where E = the elasticity, E_{pseudo} = the pseudo-elasticity, β = the coefficient for the variable of interest in the count model, and x = the variable of interest.

Table 1
Variable Descriptions.

Variable	Variable Description
data_year	Year Identifier (2005–2012)
total_crash	Total Crash Count
injury	Total Injury Crashes (Not Including Fatal Crashes)
pdo	Total Non-Injury Crashes (Property Damage Only)
Roadside Hazard Rating	Roadside Hazard Rating (1–7 where 1 is a flat, clear roadside and 7 is the highest risk roadside)
RHR45	1 = Roadside Hazard Rating is "4" or "5", 0 = Otherwise
RHR67	1 = Roadside Hazard Rating is "6" or "7", 0 = Otherwise
LNAADT	The natural log of average daily traffic volume
LNLENGTH	The natural log of segment length (miles)
Curve Density	The number of horizontal curves per mile
Average Degree of Curve	The average Degree of Curve (degrees/100 ft)
Access Density	The number of access points per mile
Passing Zone	1 = Passing Zone Present, 0 = Otherwise
Low Speed	Posted Speed 45 mph or under
Shoulder Rumble Strips	1 = Shoulder Rumble Strips Present, 0 = Otherwise

The elasticity for a binary logit is defined as shown in Eq. (13) (Train, 2009) and the pseudo-elasticity is defined as shown in Eq. (14).

$$E = \beta x(1 - P_{report}) \tag{13}$$

$$E_{pseudo} = 100\left(\frac{P_{report,with} - P_{report,without}}{P_{report,without}}\right) \tag{14}$$

Where β = the coefficient for the variable of interest in the count model, and x =the variable of interest, P_{report} = the probability that crashes are reported (as defined in Eq. (5)),$P_{report,with}$ =the probability of reporting when the variable of interest has a value of 1, and $P_{report,without}$ = the probability of reporting when the variable of interest has a value of 0.

Another important finding in recent transportation safety research is related to correlation in the error terms between different crash types. When there are correlations between different crash types (including different crash severities), estimation of separate models has been noted as being problematic due to inefficient estimates (Barua et al., 2016; Dong et al., 2014; Mannering et al., 2016; Serhiyenko et al., 2016; Zhan et al., 2015). Thus, estimation of injury and PDO models separately could potentially be problematic when determining relationships between crash frequency and site-specific safety-influencing roadway features. While this is an important concept, the purpose of this paper is to determine the plausibility of estimating underreporting models in traffic safety, and to compare underreporting model results with a common form of count regression used in traffic safety research.

In summary, the negative binomial and Poisson underreporting models are compared to a random parameters negative binomial model, which does not account for underreporting, in the present study. The data used for the analysis are described below.

4. Data

The data used in the present study included yearly data (2005–2012) from Pennsylvania and encompassed 21,340 segments of two-lane rural highways (one to three digit routes) that are owned and maintained by PennDOT. Variable definitions and descriptive statistics are shown in Tables 1 and 2, respectively. The data in Table 2 were collected as a part of a recent research project, using PennDOT electronic roadway inventory and crash data files, video photo logs, and Google Earth satellite imagery (Donnell et al., 2014).

Total crashes include all crash types and severities (PDO, injury, and fatal crashes). Injury crashes are all crashes that involve some level of injury, but no fatalities. The fatal crashes were not combined with the injury crashes in order to model the underreporting of injury crashes without confounding the results by including crashes that are highly likely to be reported. However, the fatal crashes were included in the total crash models since models for total crashes are common in traffic safety research. PDO crashes include all crashes that do not involve any of the occupants being injured as a result of the crash. The roadside hazard ratings were assigned on a 1–7 scale developed in previous research (Zeeger et al., 1987). A rating of 1 is generally considered a roadside that is free of fixed objects and contains recoverable slopes in an area that extends 30 feet or more from the outside edge of the traveled way. A rating of 7 has non-traversable roadside slopes or fixed objects in an area that is less than 5 feet from the outside edge of the traveled way. All other variables are as defined in Table 1.

5. Results and discussion

The modeling methods described in Section 3 were applied to crash data disaggregated by severity (i.e., injury and PDO crashes), and to total crashes, which included all crash types and severities combined. Models for fatal crashes failed to converge when estimating the Poisson and negative binomial underreporting models. This was not surprising since most, if not all, fatal crashes in any traffic safety data files are not subject to underreporting (Hauer and Hakkart, 1988). Thus, models for total, injury, and property damage only (PDO) crashes were estimated and are presented and discussed in this section.

As discussed in the Methodology section, Poisson underreporting, negative binomial underreporting, and random parameters negative binomial models for reported crashes were estimated. All random parameters where estimated using normal distributions. These models are shown in Tables 3–5 for total crashes, injury crashes, and PDO crashes, respectively. The elasticities and pseudo-elasticities for each of the predictor variables are also shown in Tables 3–5. Given that the signs and magnitudes of the coefficients for both Poisson and negative binomial models were consistent, and the log-likelihoods and MSE values indicated that the negative binomial underreporting models fit the data much better than the Poisson underreporting models, the remaining discussion focuses on the negative binomial underreporting models. The interpretations given for the negative binomial underreporting models can also be applied to the Poisson underreporting models.

5.1. Count model interpretations

The signs and magnitudes of all coefficients in Tables 3–5 are consistent with engineering intuition. When comparing the regression coefficients in the count portion of the negative binomial underreporting models to the coefficients in the random parameters negative binomial models, many of the variables have the same sign and similar magnitudes. In the random parameters models, the intercept, access density, curve density, and average degree of curve were all found to be random parameters. For both count models, the signs and magnitudes of the coefficients for the natural log of traffic volume and natural log of segment length are consistent with previous research (Donnell et al., 2014). The negative coefficients for the presence of a passing zone are likely the result of passing zones being present on road segments with long available sight distances, which is correlated with fewer crashes. While the ability of cars to pass is not likely to cause a reduction in crashes, the presence of

Table 2
Descriptive Statistics.

Variable	Obs	Mean	Std. Dev.	Min	Max
total_crash	170,468	0.667	1.144	0	23
injury	170,468	0.347	0.724	0	13
pdo	170,468	0.306	0.672	0	13
Roadside Hazard Rating	170,468	4.861	0.815	1	7
LNAADT	170,468	7.701	0.964	4.304	10.264
LNLENGTH	170,468	−0.794	0.347	−5.799	0.389
Curve Density	170,468	2.299	2.506	0	42.581
Average Degree of Curve	170,468	3.804	6.068	0	194.551
Access Density	170,468	16.297	14.307	0	330
Variable	Obs	Percentage With Value of 1			
Passing Zone	170,468	0.2838			
Low Speed	170,468	0.4541			
Shoulder Rumble Strips	170,468	0.0814			
RHR45	170,468	0.7468			
RHR67	170,468	0.1960			
data_year = 2005	170,468	0.2500			
data_year = 2006	170,468	0.2500			
data_year = 2007	170,468	0.2500			
data_year = 2008	170,468	0.2500			
data_year = 2009	170,468	0.2500			
data_year = 2010	170,468	0.2500			
data_year = 2011	170,468	0.2500			
data_year = 2012	170,468	0.2500			

Table 3
Total Crash Models.

Model Type	Crash Severity (Outcome)	Random Parameters			Poisson Underreporting			Negative Binomial Underreporting			Absolute Difference in Elasticities (random parameters and underreporting NB)
	Number of Segment-Years	170,468			170,468			170,468			
	Log Likelihood	−167,171.00			−173,355.10			−80,505.69			
	MSE	1.1991			1.0714			1.0406			
	Variable	Coeff.	St. Error	Elast.	Coeff.	St. Error	Elast.	Coeff.	St. Error	Elast.	
Count Model	Intercept	−6.1745	0.0236	–	***−5.9198***	***0.0679***	–	***−5.962***	***0.1019***		–
	LNAADT	***0.7364***	***0.0026***	***5.66***	***0.7599***	***0.0046***	***5.85***	***0.7581***	***0.0050***	***5.85***	***0.19***
	LNLENGTH	***0.6810***	***0.0059***	***0.54***	***0.6398***	***0.0106***	***0.51***	***0.6255***	***0.0141***	***0.50***	***0.04***
	Passing Zone	***−0.1894***	***0.0050***	***−17.25***	***−0.1469***	***0.027***	***−13.66***	***−0.0025***	***0.0007***	***−0.46***	***16.79***
	Access Density	***0.0073***	***0.0001***	***0.12***	***−0.0023***	***0.0007***	***−0.04***	***−0.0023***	***0.0007***	***−0.04***	***0.16***
	σ – Access Density	***0.0042***	***0.0001***	–	–	–	–	–	–	–	–
	Low Speed	***0.0954***	***0.0045***	***10.79***	**0.0484**	**0.0235**	**4.96**	0.0499	0.0335	4.97	5.82
	RHR45	***0.0834***	***0.0090***	***8.95***	***0.0872***	***0.016***	***9.11***	***0.0853***	***0.0165***	***9.09***	***0.14***
	RHR67	***0.0750***	***0.0101***	***8.54***	***0.0882***	***0.0179***	***9.22***	***0.0862***	***0.0183***	***9.11***	***0.57***
	Curve Density	***0.0197***	***0.0010***	***0.07***	***0.0303***	***0.0036***	***0.07***	***0.0303***	***0.0036***	***0.07***	***0.00***
	σ – Curve Density	***0.0289***	***0.0007***	–	–	–	–	–	–	–	–
	Average Degree of Curve	***0.0158***	***0.0003***	***0.04***	−0.0006	0.0011	0.002	−0.0007	0.0015	0.002	0.04
	σ – Deg. Of Curve	***0.0127***	***0.0003***	–	–	–	–	–	–	–	–
	Shoulder Rumble Strips	***−0.1396***	***0.0076***	***−12.85***	***−0.2094***	***0.0407***	***−18.89***	***−0.2236***	***0.0408***	***−2.12***	***10.73***
	α	***0.0773***	***0.0121***	–	–	–	–			–	–
	σ – Intercept	***0.5756***	***0.0019***	–	***0.6632***	***0.0047***	–	–	–	–	–
Logit Reporting Model	Intercept	–	–	–	−0.1311	0.108	–	−0.1251	0.1761	–	–
	Passing Zone	–	–	–	−0.1116	0.0744	−1.14	−0.1070	0.0813	−1.03	–
	Access Density	–	–	–	***0.0398***	***0.0028***	***0.19***	***0.0396***	***0.0059***	***0.19***	–
	Low Speed	–	–	–	**0.1592**	**0.0684**	**2.49**	0.1663	0.0906	2.62	–
	Curve Density	–	–	–	***−0.1083***	***0.014***	***−0.07***	***−0.1110***	***0.0191***	***−0.08***	–
	Average Degree of Curve	–	–	–	***0.1425***	***0.0125***	***0.16***	***0.1418***	***0.0181***	***0.15***	–
	Shoulder Rumble Strips	–	–	–	0.1888	0.1207	0.56	0.2350	0.1761	0.92	–

Note: Bold: P-value < 0.05, Bold Italicized: P-value < 0.01

a passing zone likely acts as a proxy variable for other unobserved factors that are associated with fewer crashes (e.g., flatter curves, wider shoulders, long available sight distances etc.).

The estimated coefficients for access density were found to be random in the random parameters models and indicate that crash frequency increases as access density increases. This finding indicates that more access points are associated with a larger number of conflicts that could lead to higher crash frequencies. However, the negative binomial underreporting models indicate that an increase in access points is associated with a decrease in crash frequency, although the effect is not statistically significant for injury crashes. While the decrease in crashes associated with increased access density is counterintuitive, the inclusion of access density in the reporting model may be picking up the effects of unobserved variables related to the area type (e.g., rural residential area will have a high access density).

The coefficients for low speed indicate that crash frequencies are higher on roads with posted speeds below 50 mph than on roads

Table 4
Injury Crash Models.

Model Type	Crash Severity (Outcome)	Random Parameters			Poisson Underreporting			Negative Binomial Underreporting			Absolute Difference in Elasticities (random parameters and underreporting NB)
	Number of Segment-Years	170,468			170,468			170,468			
	Log Likelihood	−117,077.30			−120,118.80			−91,340.33			
	MSE	0.5392			0.4993			0.4528			
	Variable	Coeff.	St. Error	Elast.	Coeff.	St. Error	Elast.	Coeff.	St. Error	Elast.	
Count Model	Intercept	−6.6940	0.0312	–	***−6.5821***	***0.0718***	–	***−6.3549***	***0.0768***	–	–
	LNAADT	***0.7243***	***0.0034***	***5.57***	***0.7441***	***0.0059***	***5.73***	***0.7429***	***0.0061***	***5.72***	***0.15***
	LNLENGTH	***0.6870***	***0.0081***	***0.55***	***0.6347***	***0.0142***	***0.50***	***0.6279***	***0.0155***	***0.49***	***0.06***
	Passing Zone	***−0.2000***	***0.0067***	***−18.07***	***−0.1285***	***0.0272***	***−12.06***	***−0.1295***	***0.0257***	***−12.06***	***6.01***
	Access Density	***0.0079***	***0.0002***	***0.13***	−0.0006	0.0007	−0.01	−0.0007	0.0007	−0.01	***0.14***
	σ – Access Density	***0.0033***	***0.0001***	–	–	–	–	–	–	–	–
	Low Speed	***0.1016***	***0.0060***	***11.02***	**0.0707**	**0.0227**	**7.33**	**0.0745**	**0.0238**	**7.34**	***3.68***
	RHR45	***0.0387***	***0.0118***	***4.19***	**0.0426**	**0.0204**	**4.35**	**0.0415**	**0.0202**	**4.17**	***0.02***
	RHR67	*0.0257*	*0.0133*	***3.07***	0.0362	0.0229	3.69	0.0350	0.0227	3.65	***0.58***
	Curve Density	***0.0219***	***0.0014***	***0.07***	***0.0315***	***0.0038***	***0.07***	***0.0319***	***0.0035***	***0.08***	***0.01***
	σ – Curve Density	***0.0325***	***0.0009***	–	–	–	–	–	–	–	–
	Average Degree of Curve	***0.0132***	***0.0005***	***0.04***	0.0011	0.0012	0.11	0.0011	0.0013	0.11	***0.07***
	σ – Deg. Of Curve	***0.0097***	***0.0004***	–	–	–	–	–	–	–	–
	Shoulder Rumble Strips	***−0.1494***	***0.0100***	***−13.86***	***−0.2062***	***0.0386***	***−18.63***	***−0.2104***	***0.0384***	***−18.98***	***5.12***
	α	***0.0719***	***0.0042***	–	–	–	–	***0.5267***	***0.0126***	–	–
	σ – Intercept	***0.6051***	***0.0026***	–	***0.6749***	***0.0071***	–	–	–	–	–
Logit Reporting Model	Intercept	–	–	–	0.1035	0.116	–	0.1160	0.1263	–	–
	Passing Zone	–	–	–	**−0.2315**	**0.0933**	**−1.82**	**−0.2289**	**0.0930**	**−1.79**	–
	Access Density	–	–	–	***0.0565***	***0.0052***	***0.41***	***0.0571***	***0.0072***	***0.40***	–
	Low Speed	–	–	–	0.1194	0.0868	1.32	0.1162	0.0918	1.30	–
	Curve Density	–	–	–	***−0.1515***	***0.0208***	***−0.15***	***−0.1542***	***0.0215***	***−0.15***	–
	Average Degree of Curve	–	–	–	***0.1879***	***0.0219***	***0.32***	***0.1886***	***0.0225***	***0.33***	–
	Shoulder Rumble Strips	–	–	–	0.2308	0.1481	0.53	0.2608	0.1549	0.56	–

Note: Bold: P-value < 0.05, Bold Italicized: P-value < 0.01

Table 5
PDO Crash Models.

Model Type	Crash Severity (Outcome)	Random Parameters			Poisson Underreporting			Negative Binomial Underreporting			Absolute Difference in Elasticities (random parameters and underreporting NB)
	Number of Segment-Years	170,468			170,468			170,468			
	Log Likelihood	−108507.40			−110,926.70			−87,423.43			
	MSE	0.4201			0.4010			0.3936			
	Variable	Coeff.	St. Error	Elast.	Coeff.	St. Error	Elast.	Coeff.	St. Error	Elast.	
Count Model	Intercept	−7.2461	0.0342	–	***−6.6443***	***0.1417***	–	***−6.3982***	***0.2025***	–	–
	LNAADT	***0.7727***	***0.0037***	***5.94***	***0.7898***	***0.0065***	***6.08***	***0.7888***	***0.0066***	***6.07***	0.13
	LNLENGTH	***0.6869***	***0.0083***	***0.55***	***0.6481***	***0.0147***	***0.51***	***0.6419***	***0.0192***	***0.50***	0.05
	Passing Zone	***−0.1892***	***0.0072***	***−17.41***	***−0.2491***	***0.0488***	***−22.05***	***−0.2519***	***0.0622***	***−22.27***	4.86
	Access Density	***0.0055***	***0.0002***	***0.09***	***−0.0044***	***0.0011***	***−0.07***	***−0.0045***	***0.0011***	***−0.07***	0.16
	σ – Access Density	***0.0025***	***0.0001***	–	–	–	–	–	–	–	–
	Low Speed	***0.1152***	***0.0064***	***12.48***	−0.0296	0.0511	−2.92	−0.0323	0.0822	−3.21	15.69
	RHR45	***0.1312***	***0.0131***	***13.79***	***0.1292***	***0.0225***	***13.79***	***0.1301***	***0.0221***	***13.79***	0.00
	RHR67	***0.1311***	***0.0145***	***13.96***	***0.1304***	***0.0250***	***13.93***	***0.1311***	***0.0247***	***13.96***	0.00
	Curve Density	***0.0166***	***0.0015***	***0.07***	***0.0273***	***0.0069***	***0.06***	***0.0282***	***0.0054***	***0.07***	0.00
	σ – Curve Density	***0.0461***	***0.0010***	–	–	–	–	–	–	–	–
	Average Degree of Curve	***0.0150***	***0.0005***	***0.04***	**−0.0041**	**0.0019**	**−0.02**	**−0.0042**	0.0024	−0.02	0.06
	σ – Deg. Of Curve	***0.0132***	***0.0004***	–	–	–	–	–	–	–	–
	Shoulder Rumble Strips	***−0.1537***	***0.0109***	***−13.78***	**−0.1967**	**0.0797**	**−17.86**	**−0.2076**	**0.0885**	**−18.89**	5.11
	α	***0.1067***	***0.0051***	–	–	–	–	***0.5588***	***0.0150***	–	–
	σ – Intercept	***0.5826***	***0.0029***	–	***0.6885***	***0.0079***	–	–	–	–	–
Logit Reporting Model	Intercept	–	–	–	***−0.6481***	***0.1986***	–	**−0.6574**	**0.3006**	–	–
	Passing Zone	–	–	–	0.1391	0.1042	−2.01	0.1432	0.1252	−2.09	–
	Access Density	–	–	–	***0.0260***	***0.0027***	***0.27***	**0.0259**	**0.0042**	**0.27**	–
	Low Speed	–	–	–	**0.3109**	**0.102**	**7.89**	**0.3183**	**0.1494**	**7.93**	–
	Curve Density	–	–	–	***−0.0638***	***0.0178***	***−0.09***	***−0.0646***	***0.0195***	***−0.09***	–
	Average Degree of Curve	–	–	–	***0.1004***	***0.0114***	***0.25***	***0.0999***	***0.0143***	***0.25***	–
	Shoulder Rumble Strips	–	–	–	0.0907	0.1671	3.90	0.1160	0.1939	5.7	–

Note: Bold: P-value < 0.05, Bold Italicized: P-value < 0.01

with posted speeds 50 mph or higher. The only exception is for the PDO crashes using the negative binomial underreporting model, which found that the effect was not statistically significant. These findings are consistent with previous research (Butsick et al., 2015; Malshkina and Mannering, 2010; Wood and Donnell, 2016; Wood and Porter, 2013; Wood et al., 2015a,b).

The positive coefficients for the roadside hazard rating indicator variables, relative to the baseline roadside hazard ratings of 1–3, are consistent with engineering intuition. It is expected that as the roadside hazard rating increases, the crash frequency would increase. This is confirmed by the findings using both regression methods and is consistent with established safety effects for roadside hazard rating (AASHTO, 2010).

The positive coefficients for curve density indicate that crash frequency increases as the number of horizontal curves along the road segment increases. The coefficient was found to be random in each of the random parameters models. This finding is consistent with previous research (Butsick et al., 2015; Kweon and Oh, 2011).

The average degree of curve is proportional to the inverse of curve radius (i.e., degree of curve = 5729.58/Radius). Thus, positive coefficients for the average degree of curve indicate that crash frequency increases as the radius of curve decreases while negative coefficients have the opposite interpretation. The coefficients for average degree of curve in the random parameters models are all positive (and random), as is the coefficient for injury crashes in the negative binomial underreporting model. The negative coefficients in the total and PDO crash negative binomial underreporting models (crash frequency portion) are counterintuitive, although only the coefficient in the PDO model is statistically significant and the magnitudes of the coefficients are negligible. However, the very small negative coefficients in the frequency model indicate that the effect of the average degree of curve does not have a practical effect on the true crash frequency (0.01–0.4% change). The change in reported crashes is due to the change in the probability for reporting crashes (i.e., increased probability leading to an increase in reported crashes as indicated by the logit model coefficients). It is likely that sharp horizontal curves lead to increased vehicle damage and a possible increase in crash severity as a result of increased roadway departure angles. The increased angles for run-off-road crashes (and thus, increased vehicle damage, etc.) leads to an increased probability of crash reporting.

Previous research has found that shoulder rumble strips are associated with reduced crash frequencies (Torbic et al., 2009), thus the negative coefficients for shoulder rumble strip presence in the underreporting and random parameters negative binomial models are consistent with past research.

5.2. Reporting model interpretations

Interpretation of the reporting models requires understanding the relationships between the variables included and the probability of reporting. The sign and magnitude of the intercepts is also of interest. In the reporting models for total and PDO crashes, the intercepts have negative signs. The intercept for the injury reporting model is positive. The signs and magnitude of the intercepts are consistent with expectation for all of the reporting models. The negative signs are for the models with a higher probability of underreporting, which includes the lowest severity level (i.e., PDO) and total crashes. With regards to total crashes, it is worth noting that PDO crashes represent nearly 31 percent of total crashes, so it was anticipated that these two models would produce similar signs in the models estimating reporting probabilities. The positive intercept for injury crash reporting indicates that the probability of reporting injury crashes is much larger than for total or PDO crashes.

The coefficients for passing zones indicate that the probability of reporting total and injury crashes is lower on road segments with passing zones, but the probability of reporting PDO crashes is higher on road segments with passing zones. As with the count model, this may be due to the passing zone indicator variable acting as a proxy variable for other factors that affect reporting.

The estimated effect of increased access density on crash reporting is an increase in the probability of reporting total, injury, and PDO crashes. This may be picking up the effects of unobserved variables that are correlated with increased access density.

Low speed roads were found to have higher probabilities of crash reporting than high speed roads for total, injury, and PDO crashes. This may be a contributing factor to the magnitude of findings in previous research (Donnell et al., 2010; Malshkina and Mannering, 2010; Wood and Porter, 2013; Wood et al., 2015a,b), which shows that low speed roads have higher reported crash frequencies, even when controlling for many potential confounding factors.

When curve density increases, the probability of reporting crashes decreases for total, injury, and PDO crashes. While the reason for this is not clear, curve density could be associated with multiple factors that influence crash underreporting such as crash severity, the level of perceived damage, and police willingness to report the crash.

Based on the coefficients for the average degree of curve in the reporting model, the probability of crashes being reported increases as the radius of curve decreases (or degree of curve increases). This is likely due to an increase in run-off road crashes and an increase in crash severity.

The increase in the probability of reporting associated with the presence of shoulder rumble strips was not anticipated. For total, injury, and PDO crashes, the probability of reporting increases with the presence of rumble strips, but the effect is not statistically significant. This increase may be the result of a confounding factor and suggests that the safety effects of shoulder rumble strips might be reconsidered in future research.

5.3. Model comparisons

To compare the elasticities for reported crashes estimated using the negative binomial underreporting and random parameters negative binomial models, the differences between the two model types were calculated and are provided in Tables 3–5.

The differences in the elasticities between the two model types are noteworthy because the differences are much greater for the indicator variables (pseudo-elasticities) than for the continuous variables. The indicator variables that were found to influence the probability of reporting (passing zone, access density, low speed, curve density, average degree of curvature, and centerline rumble strips) all had absolute differences in pseudo-elasticities of 0.00–16.79%, which indicates significant differences in the estimates between the negative binomial underreporting and random parameters negative binomial estimates for these variables. These differences highlight the issue of endogeneity bias, which occurs when the reported crashes are correlated with underreporting, which is subsequently not considered in a count regression model. Since this is the case, the negative binomial underreporting model results are more consistent with the theory of how crash reporting occurs and are thus less likely to yield biased CMFs than when using models that do not account for underreporting.

The differences for the continuous variables and the indicator variables that did not influence the probability of reporting were much smaller, with absolute magnitudes in the range of 0.00–0.58%. However, it should be noted that, with the exception of the elasticities for traffic volumes, the magnitudes for all continuous variables were very small, leading to small differences in the elasticities for the two regression methods. Thus, the differences in the estimated

Table 6
PDO Negative Binomial Underreporting Model Validation.

Miles	Crash Type	2010	2011	2012	2013	2014	Total
7.173	Reported	7	6	5	6	7	31
	Unreported	5	6	8	6	5	30
	Prob. Reporting	0.567					
	Predicted Unreported	5.35	4.58	3.82	4.58	5.35	23.7
	Predicted – Observed Underreported	0.35	−1.42	−4.18	−1.42	0.35	−6.33
4.288	Reported	2	5	3	2	5	17
	Unreported	2	4	0	6	1	13
	Prob. Reporting	0.531					
	Predicted Unreported	1.77	4.42	2.65	1.77	4.42	15.01
	Predicted – Observed Underreported	−0.23	0.42	2.65	−4.23	3.42	2.01
4.306	Reported	3	1	2	3	6	15
	Unreported	1	0	1	1	2	5
	Prob. Reporting	0.600					
	Predicted Unreported	2.00	0.67	1.33	2.00	4.00	10.00
	Predicted – Observed Underreported	1.00	0.67	0.33	1.00	2.00	5.00
1.159	Reported	2	0	1	2	1	6
	Unreported	3	0	1	3	1	8
	Prob. Reporting	0.516					
	Predicted Unreported	1.87	0.00	0.94	1.87	0.94	5.62
	Predicted – Observed Underreported	−1.13	0.00	−0.06	−1.13	−0.06	−2.38
3.559	Reported	3	3	2	3	3	14
	Unreported	3	2	1	2	2	10
	Prob. Reporting	0.567					
	Predicted Unreported	2.29	2.29	1.52	2.29	2.29	10.67
	Predicted – Observed Underreported	−0.71	0.29	0.52	0.29	0.29	0.67
2.213	Reported	5	4	0	5	9	23
	Unreported	6	0	1	4	2	13
	Prob. Reporting	0.523					
	Predicted Unreported	4.56	3.65	0.00	4.56	8.21	20.97
	Predicted – Observed Underreported	−1.44	3.65	−1.00	0.56	6.21	7.97
2.989	Reported	0	5	1	4	5	15
	Unreported	2	2	2	2	1	9
	Prob. Reporting	0.540					
	Predicted Unreported	0.00	4.26	0.85	3.41	4.26	12.78
	Predicted – Observed Underreported	−2.00	2.26	−1.15	1.41	3.26	3.78
2.314	Reported	0	2	2	3	4	11
	Unreported	3	1	1	1	1	7
	Prob. Reporting	0.492					
	Predicted Unreported	0.00	2.06	2.06	3.09	4.12	11.34
	Predicted – Observed Underreported	−3.00	1.06	1.06	2.09	3.12	4.34
Total Predicted Underreported Crashes						110.07	
Total Underreported Crashes						95	
Total Miles of Roadway						27.912	

elasticities, while small, are still likely to be biased when they influence the probability of reporting and are not considered in crash frequency regression models.

If a variable does not affect the probability of reporting, the elasticities for reported crashes and the true crashes are essentially the same due to no endogeneity between the variable effect on crash frequency and the probability of crashes being reported. Thus, if it is known that a variable does not affect the probability of crash reporting, traditional count regression or before-after analysis methods may be used to develop CMFs. However, since the relationships between design elements and countermeasures with the probability of crash reporting are not well understood, the best practice would be to use underreporting models to assess the assumption of the variable not affecting the probability of crash reporting whenever possible.

Also of interest is that the mean squared error for the negative binomial underreporting models are smaller than the random parameters negative binomial models, indicating that the negative binomial underreporting models fit the data better. Thus, negative binomial underreporting models are likely less biased (in terms of predictive validity) and fit the data better than random parameters regression models.

5.4. Underreporting model validation

Given that the underreporting models attempt to account for the latent crash reporting process, a validation exercise was performed. Three local townships in Pennsylvania provided data for crashes that police responded to, but where the crash was not considered reportable to PennDOT (always PDO), for a total of eight two-lane rural roads that are owned and maintained by PennDOT. Thus, these crashes encompass some, but not likely all, non-reportable crashes for these roads. These non-reported crashes were compared with the observed PDO crashes and the predicted probabilities for crash reporting from the negative binomial underreporting model. These comparisons are shown in Table 6.

As indicated in Table 6, there were 110.07 non-reportable total crashes predicted for these roads between the years 2010 and

2014 (5 years). The total observed non-reportable crashes was 95 crashes for the same period on the same roads. Given that the observed non-reportable crashes are likely a subset of the actual number of non-reportable crashes, this difference (15.07 crashes) is reasonable and provides preliminary evidence that the negative binomial underreporting models can be used to develop predictive models that provide better predictions for observed crashes (i.e., improved MSE values from Tables 3–5) as well as providing reasonable models for predicting the probability that crashes were actually reported. This would provide engineers with the ability to predict the observed number of crashes, unreported number of crashes, and total crashes for different conditions.

6. Conclusions and recommendations

This study described why it is necessary to account for underreporting when estimating the effects of a variable on expected crash frequency. When a variable affects both the probability of crash reporting and the true crash frequency, failure to account for and model the probability of reporting leads to biased inferences. Several research articles in the social sciences and epidemiology have also shown that not accounting for underreporting or systematic differences in reporting in different contexts leads to biased inference (Davenport, 2009; Dvorzak and Wagner, 2015; Earl et al., 2004; Fariss, 2015, 2014; Hill et al., 2013; Pararai et al., 2010, 2006; Schnakenberg and Fariss, 2014).

Elasticities for variables estimated using the Poisson underreporting models with heterogeneity, negative binomial underreporting models, and the random parameters negative binomial regression were estimated and compared in the present study. The data used included yearly observations for two-lane rural highway segments in Pennsylvania. The elasticity comparisons from the two methods found that, consistent with theory, if the variable impacts both the probability of crashes being reported and the true crash frequency, regression methods that do not model and account for underreporting yield biased estimates due to endogeneity. The magnitude of the bias due to not accounting for underreporting was found to be much larger for indicator variables than for the continuous variables. If a variable is related to the true crash frequency and not to the probability of crashes being reported, the estimated elasticities using random parameters models are consistent with the results of the negative binomial underreporting models. For the binary indicator variables considered in this study, endogeneity produced elasticity differences as high as 16.79%. For the variables that were only associated with the change in the true crash frequency, the differences in the estimated elasticities were negligible.

It was also found that the negative binomial underreporting models provided better predictions than the random parameters models (as indicated by the mean squared error and log-likelihood values). Thus, negative binomial underreporting models may be beneficial to implement when developing prediction models in future crash frequency research. While it is possible that negative binomial underreporting models could be used in empirical Bayes before-after studies, using the empirical Bayes before-after method requires the assumption that the treatment does not affect the probability of crash reporting. This assumption is implicitly made regardless of the regression method used to estimate the SPF and for all before-after analytical methods. Thus, any before-after studies that estimate CMFs may result in biased estimates if the treatment is correlated with the probability of crash reporting.

The comparison of predicted underreported crashes with crashes that were reported to police, but were not included in the PennDOT crash database provided preliminary evidence that the negative binomial underreporting models can provide reasonable predictions of the probability of reporting (or underreporting). This could prove to be valuable for engineers working with safety prediction models as this would give them the ability to predict the reported, non-reportable, and total crashes.

Since there are many paths that can link variables in crash frequency research with the probability of crash reporting (e.g., crash severity distribution, distribution of crash costs, and the willingness of drivers to report crashes to police), it would be beneficial for researchers to use negative binomial underreporting models to assess whether the variables of interest impact both the probability of crash reporting and the true crash frequency, or just one of them. If any of the variables of interest impact both, then the negative binomial underreporting model should be used. If it only impacts the true crash frequency, other analysis methods such as the random parameters negative binomial may be considered.

Future research should consider the use of Bayesian underreporting models. Methods for evaluating the impacts of underreporting in a Bayesian framework have the potential to provide insights into sources of heterogeneity that can be explained using random parameters. The Bayesian framework more generally provides a principled set of methods for model development and model comparison that can begin with simple models and then expand into more complex alternatives (Blei, 2014; Gelman and Shalizi, 2013; Schnakenberg and Fariss, 2014).

Another avenue of future research that could have important theoretical implications is related to bivariate/multivariate formulations of underreporting models. In this framework, the bivariate models could be used to jointly estimate injury (including fatal crashes) and PDO crashes, while multivariate models could jointly estimate fatal, injury, and PDO crashes. This paper considered the frequency of each severity type separately. There have recently been multiple applications of random parameters multivariate count models in the traffic safety literature, each of which has shown that it is important to account for correlation between different crash types and regression coefficients (Barua et al., 2016; Dong et al., 2014; Serhiyenko et al., 2016; Zhan et al., 2015). Comparing Bayesian underreporting models with multivariate random parameters models, as well as developing bivariate/multivariate underreporting models could lead to important research advances.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aap.2016.06.013.

References

AASHTO, 2010. Highway Safety Manual. American Association of State Highway and Transportation Officials, Washington, D.C.
Abay, K.A., 2015. Investigating the nature and impact of reporting bias in road crash data. Transp. Res. Part A Policy Pract. 71, 31–45.
Alsop, J., Langley, J., 2001. Under-Reporting of motor vehicle traffic crash victims in New Zealand. Accid. Anal. Prev. 33, 353–359.
Amoros, E., Martin, J.-L., Laumon, B., 2006. Under-reporting of road crash casualties in France. Accid. Anal. Prev. 38, 627–635.
Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Anal. Methods Accid. Res. 9, 1–15.
Blei, D.M., 2014. Build, compute, critique, repeat: data analysis with latent variable models. Annu. Rev. Stat. Appl. 1, 203–232.
Brookoff, D., Campbell, E.A., Shaw, L.M., 1993. The underreporting of cocaine-related trauma: drug abuse warning network reports vs hospital toxicology tests. Am. J. Public Health 83, 369–371.
Butsick, A.J., Jovanis, P.P., Wood, J.S., 2015. Modeling safety effects of geometric design consistency on two-lane rural roads using mixed effects negative binomial regression. Transp. Res. Board 94th Annu. Meet., 15–0797.
Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press.
Davenport, C., 2009. Media Bias, Perspective, and State Repression: The Black Panther Party. Cambridge University Press.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections. Accid. Anal. Prev. 70, 320–329.
Donnell, E.T., Porter, R.J., Shankar, V.N., 2010. A framework for estimating the safety effects of roadway lighting at intersections. Saf. Sci. 48, 1436–1444.
Donnell, E.T., Gayah, V., Jovanis, P.P., 2014. Safety performance functions: final report. Pennsylvania Dept. Transp., Rep. No. FHWA-PA-2014-007-PSU WO 1.
Dvorzak, M., Wagner, H., 2015. Sparse bayesian modelling of underreported count data. Stat. Model.
Earl, J., Martin, A., McCarthy, J.D., Soule, S.A., 2004. The use of newspaper data in the study of collective action. Annu. Rev. Sociol. 30, 65–80.
Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. Transp. Res. Rec. J. Transp. Res. Board 1665, 133–140.
Fariss, C.J., 2014. Respect for human rights has improved over time: modeling the changing standard of accountability. Am. Polit. Sci. Rev. 108, 297–318.
Fariss, C.J., 2015. Uncertain events: a dynamic latent variable model of human rights respect and government killing with binary ordered, and count outcomes. Work. Pap.
Gelman, A., Shalizi, C.R., 2013. Philosophy and the practice of bayesian statistics. Br. J. Math. Stat. Psychol. 66, 8–38.
Greene, W.H., 2007. Limdep Econometric Modeling Guide Version 9.0.
Greene, W.H., 2011. Econometric Analysis, 7th ed. Prentice Hall.
Hauer, E., Hakkart, A., 1988. Extent and some implications of incomplete accident reporting. Transp. Res. Rec. J. Transp. Res. Board 1185, 1–10.
Hauer, E., 2006. The frequency-severity indeterminacy. Accid. Anal. Prev. 38, 78–83.
Hilbe, J.M., 2011. Negative Binomial Regression. Cambridge University Press.
Hill, D.W., Moore, W.H., Mukherjee, B., 2013. Information politics versus organizational incentives: when are amnesty international's 'Naming and shaming' reports biased? Int. Stud. Q. 57, 219–232.
Hosios, A.J., Peters, M., 1989. Repeated insurance contracts with adverse selection and limited commitment. Q. J. Econ. 104, 229–253.
Kamura, S.S.P., Chin, H.C., 2005. Application of poisson underreporting model to examine crash frequencies at signalized three-legged intersections. Transp. Res. Rec. J. Transp. Res. Board 1908, 46–50.
Kemp, C.D., 1973. Research note: an elementry ambiguity in accident theory. Accid. Anal. Prev. 5, 371–373.
Kennedy, P., 2008. A Guide to Econometrics, 6th ed. Blackwell Publishing, Malden, MA.
Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accid. Anal. Prev. 34, 313–321.
Kweon, Y.J., Oh, C., 2011. Identifying promising highway segments for safety improvement through speed management. Transp. Res. Rec. J. Transp. Res. Board 2213, 46–52.
Leigh, J.P., Marcin, J.P., Miller, T.R., 2004. An estimate of the U.S government's undercount of nonfatal occupational injuries. J. Occup. Environ. Med. 46, 10–18.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44, 291–305.
Ma, J., Kockelman, K.M., 2006. Bayesian multivariate poisson regression for models of injury count, by severity. Transp. Res. Rec. J. Transp. Res. Board 1950, 24–34.
Ma, J., Li, Z., 2010. Bayesian modeling of frequency-severity indeterminacy with an application to traffic crashes on two-lane highways. ICCTP 2010: Integrated Transportation Systems, 1022–1033.
Malshkina, N., Mannering, F., 2010. Emperical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accid. Anal. Prev. 42, 131–139.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22.
Mannering, F.L., Shankar, V.N., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.
National Highway Traffic Safety Administration, 2014. Fatality Analysis Reporting System (FARS) Analytic User's Manual 1975–2012. National Highway Traffic Safety Administration, Washington, D.C.
Pararai, M., Famoye, F., Lee, C., 2006. Generalized poisson regression model for underreported counts. J. Adv. Appl. Stat. 6, 305–322.
Pararai, M., Famoye, F., Lee, C., 2010. Generalized poisson-poisson mixture model for misreported counts with an application to smoking data. J. Data Sci. 8, 607–617.
Patil, S., Geedipally, S.R., Lord, D., 2012. Analysis of crash severities using nested logit model-accounting for the underreporting of crashes. Accid. Anal. Prev. 45, 646–653.
Probst, T.M., Estrada, A.X., 2010. Accident under-reporting among employees: testing the moderating influence of psychological safety climate and supervisor enforcement of safety practices. Accid. Anal. Prev. 42, 1438–1444.
Probst, T.M., Graso, M., 2013. Pressure to produce = pressure to reduce accident reporting? Accid. Anal. Prev. 59, 580–587.
Probst, T.M., Barbaranelli, C., Petitta, L., 2013. The relationship between job insecurity and accident under-reporting: a test in two countries. Work Stress 27, 383–402.
Quddus, M.A., Wang, C., Ison, S.G., 2010. Road traffic congestion and crash severity: econometric analysis using ordered response models. J. Transp. Eng. 136, 424–435.
Rosman, D.L., Knuiman, M.W., 1994. A comparison of hospital and police road injury data. Accid. Anal. Prev. 26, 215–222.
Schnakenberg, K.E., Fariss, C.J., 2014. Dynamic patterns of human rights practices. Polit. Sci. Res. Methods 2, 1–31.
Serhiyenko, V., Mamun, S.A., Ivan, J.N., Ravishanker, N., 2016. Fast bayesian inference for modeling multivariate crash counts. Anal. Methods Accid. Res. 9, 44–53.
Torbic, D.J., Hutton, J.M., Bokenkroger, C.D., Bauer, K.M., Harwood, D.W., Gilmore, D.K., Dunn, D.K., Ronchetto, J.J., Donnell, E.T., Sommer III, H.J., Garvey, P., Persaud, B., Lyon, C., 2009. NCHRP Report 641: Guidance for the Design and Application of Shoulder and Centerline Rumble Strips.
Train, K.E., 2009. Discrete Choice Methods with Simulation. Cambridge University Press.
Winkelmann, R., 2013. Econometric Analysis of Count Data. Springer Science & Business Media.
Wood, J.S., Donnell, E.T., 2016. Safety evaluation of gontinuous green T intersections: a propensity scores-genetic matching-potential outcomes approach. Accid. Anal. Prev. 93, 1–13.
Wood, J.S., Porter, R.J., 2013. Safety impacts of design exceptions on nonfreeway segments. Transp. Res. Rec. J. Transp. Res. Board 2358, 29–37.
Wood, J.S., Donnell, E.T., Porter, R.J., 2015a. Comparison of safety effect estimates obtained from empirical bayes before-after study propensity scores-potential outcomes framework, and regression model with cross-Sectional data. Accid. Anal. Prev. 75, 144–154.
Wood, J.S., Gooch, J.P., Donnell, E.T., 2015b. Estimating the safety effects of lane widths on urban streets in nebraska using the propensity scores-potential outcomes framework. Accid. Anal. Prev. 82, 180–191.
Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accid. Anal. Prev. 40, 1320–1329.
Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accid. Anal. Prev. 59, 506–521.
Ye, F., Lord, D., 2006. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models. Transp. Res. Rec. J. Transp. Res. Board 2241, 51–58.
Zeeger, C.V., Reinfurt, D.W., Hummer, J., Herf, L., Hunter, W., 1987. Safety effects of cross-section design for two-lane roads. Fed. Highw. Adm., Report No. 1.
Zhan, X., Abdul Aziz, H.M., Ukkusuri, S.V., 2015. An efficient parallel sampling technique for multivariate poisson-lognormal model: analysis with two crash count datasets. Anal. Methods Accid.