CHAPTER

3

Crash–frequency modeling

3.1 Introduction

As described in Chapter 2—*Fundamentals and Data Collection*, crash counts are nonnegative and discrete events occurring on roadways, which require specific analysis tools for assessing the safety performance of entities, such as roadway segments, intersections, and corridors among others. Although many analysis tools have been developed and applied in a wide variety of areas, the focus of this chapter is placed on statistical models tailored specifically for the characteristics associated with safety data. A comprehensive list of models that have been used in highway safety can be found in Appendix B of this book.

The intent of this chapter is to describe the fundamental concepts related to count-data models that can be used for analyzing crash data. The chapter describes the basic nomenclature of the models that have been proposed for analyzing highway safety data and different applications of crash-frequency models (note: the models are not limited to crash data and can also be used or estimated for many types of safety data such as vehicle occupants, nonmotorized road users or traffic conflicts). These applications are dependent on the objectives of the study or the type of analysis conducted by the safety analyst. The study objectives will have a large influence on the selection of the model, which is based on the "goodness-of-fit" or "goodness-of-logic" sought by the safety analyst (Miaou and Lord, 2003). In Chapter 2, a four-stage modeling process is presented for all the model applications described later. This chapter describes the sources of data dispersion, which influences the selection of the crash-frequency model and covers key models that have been proposed for analyzing count data, along with the important or relevant information about their characteristics. The models are grouped by their intended use for handling specific characteristics associated with safety data. The chapter ends with a discussion about the modeling selection process.

https://doi.org/10.1016/B978-0-12-816818-9.00007-X

3.2 Basic nomenclature

The basic nomenclature for the models described later follows this general form:

$$y_i = f(\mathbf{x}_i'\boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{3.1}$$

where y_i is the response variable for observation i; $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of estimable parameters; $\mathbf{x}'$ is a vector of explanatory variables; and, p is the number of parameters in the model (less the intercept). As we are interested in estimating the long-term mean of the observation (i.e., mean parameterization), Eq. (3.1) can be rewritten by calculating the expected value of the response variable as follows:

$$E[y_i|\mathbf{x}_i] = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{3.2}$$

where $E[y_i|\mathbf{x}_i]$ is the expected value of the response variable (μ_i is the mean of the response variable).

Based on the generalized linear modeling relationship with an exponential canonical link function (McCullagh and Nelder, 1989), Eq. (3.2) leads to the following form:

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}). \tag{3.3}$$

All the models described later follows the form described in Eq. (3.3). Note that different functional forms exist and could be used for analyzing safety data, many of which are described in Chapter 6—Cross-Sectional and Panel Studies in Safety.

3.3 Applications of crash-frequency models

Crash-frequency models are needed as crash data are generally random and independent events. In other words, in general terms, a crash is not directly correlated to another crash in space and time. It is theoretically possible, however, that a crash that occurred say on an urban freeway section at 4:00 p.m. could directly have contributed to another crash that happened at 4:06 p.m. about 500 m upstream from the location of the first crash, also known as the secondary crash (because of the queue that build-up following the initial collision) (Chapter 10—*Capacity, Mobility and Safety* covers microscopic crash risk in greater details). However, those events are extremely rare compared to all typical events and the correlation between the two would be difficult to quantify and their influence on the model performance will most likely be negligible. Hence, the assumption of independence between observations, as defined earlier, will be the core attribute for the models described in this chapter and the rest of the textbook.

Crash-frequency (and crash-severity) models can be used for analyzing different aspects of the transportation system. The type of applications will also govern which model to select and how to evaluate the performance of different components of the system and/or its users. Later, we describe how these models can be used for this purpose.

3.3.1 Understanding relationships

The first application consists of developing statistical models with the objective of learning a useful relationship between observable features and a potentially actionable outcome about the system from which the data are extracted and analyzed. More specifically, the objective consists of establishing a relationship between safety, being crashes, vehicle occupants, and so forth, and variables that are assumed to influence safety or crash risk. In statistical terms, a relationship is established between a dependent variable and a series of independent variables. Examining the sign of a coefficient is an example of such application. In general, this application requires very tight statistical controls so that the effects of covariates are in fact independent and do reflect the attributes of the "system." Multicollinearity among variables is an enemy of this application, as much focus is on the effect of predictor variables and their interpretability. Similarly, the omitted-variable bias that arises due to a model that does not include important variables can also negatively affect the modeling results (Lord and Mannering, 2010).

3.3.2 Screening variables

The second application consists of developing models for screening purposes, where the objective is to determine which covariates or independent variables have specific or significant effects on the risk of collisions (as associative relationship, as opposed to causal relationship described later). For this application, most of the attention is devoted to the covariates of the statistical models. Usually, this is accomplished by examining the statistical significance of the variables investigated, such as at the 5%, 10%, or 15% level (i.e., p-value) of Feature Importance Score (i.e., a statistical technique used to rank the importance of each variable). In addition, multicollinearity can be problematic for this application.

3.3.3 Sensitivity of variables

The third application seeks to examine the sensitivity of the variables that have been identified as part of the screening process described earlier. Under this application, the coefficients of the model are investigated in

greater detail. This can be accomplished by examining the marginal effects of the coefficients. They are obtained by taking the partial derivative of the coefficients under investigation, one at the time. For example, the effect of changes in variable x_j (the jth coefficient) can be estimated as follows:

$$ME = \frac{\partial \mu}{\partial x_j} = \beta_j \exp(\mathbf{x}'\boldsymbol{\beta}). \tag{3.4}$$

Other approaches for investigating the marginal effects of coefficients can be found in Cameron and Trivedi (2013). An emerging trend in the Machine Learning space is to use counterfactuals to provide explanations (Molnar, 2020).

3.3.4 Prediction

The fourth application aims at developing models for prediction purposes. In this application, the goal is to develop models with the best predictive capabilities. These models could be used with the data collected as part of the model development or with a completely new dataset. Predictive models are very useful for identifying hot spots or high-risk sites (discussed in Chapter 8—*Identification of Hazardous Sites*)—as they can guide researchers in forecasting the expected number of crashes at selected locations—which are then compared to the observed values. Multicollinearity among variables is more acceptable in models aimed at prediction, although they still need to be investigated to estimate the magnitude of the correlation, as most of the focus of the overall analysis is on the dependent variable. Variance-inflation-factors is a useful tool used for identifying highly correlated variables (Myers, 2000). Note that a proper or logical relationship[1] must exist between the dependent and independent variables if the model is developed for predicting crashes or other types of safety data. This is to avoid having the model work as a black box, which could lead to biased estimates or large errors.

3.3.5 Causal relationships

As discussed in the previous sections, the statistical models are used to describe relationships between variables. In a perfect world, one would seek to establish causal relationships, meaning that being able to capture changes in risk factors that can directly influence the risk for a crash to

[1] The authors have noted models published in reports or manuscripts in which the relationship between the number of crashes and traffic flow had an inverse relationship, which meant that the crash risk tended towards infinity as the traffic flow dropped to zero and became lower as traffic flow increased (i.e., risk = 1 / flow).

occur or observe a change in the severity outcome when a crash happens. In other words, the causal relationship seeks to determine the deterministic mechanism that leads to a crash rather than rely on the statistical nature of the crash process. Although some studies have been conducted in highway safety on this topic, the field of research is still in its infancy, which can be attributed to the complexity of the crash process and the lack of detailed information related to crash occurrences found in police reports and other sources of information. Curious readers are referred to Davis (2000), Elvik (2003, 2011), and Hauer (2010, 2020) for further discussions about cause–effect relationships in highway safety. Pearl (2009) provided pioneering work in this area, which could be of interest for some readers.

3.4 Sources of dispersion

This section describes the characteristics of dispersion (both over- and underdispersion) in crash data. Overdispersion refers to the attribute that the variance of the observed or response variables is greater than the mean, $\text{Var}[y_i] > E[y_i]$. On the other hand, underdispersion refers to the variance that is smaller than the mean, $\text{Var}[y_i] < E[y_i]$. Equi-dispersion refers to when both the mean and variance are equal. With crash data, observed equi-dispersion is extremely rare (Lord and Mannering, 2010). Fig. 3.1 shows an example of the residuals (after the model is fitted to the data) that exhibit over- and underdispersion.

3.4.1 Overdispersion

As discussed in the previous chapter, the fundamental crash process shows that crash counts are usually overdispersed, which, at its core, is attributed to the nature of the data. As such, the process, when looking at the raw count data, follows a Bernoulli trial with unequal probability of independent events, which is also known as Poisson trials (Lord et al., 2005).

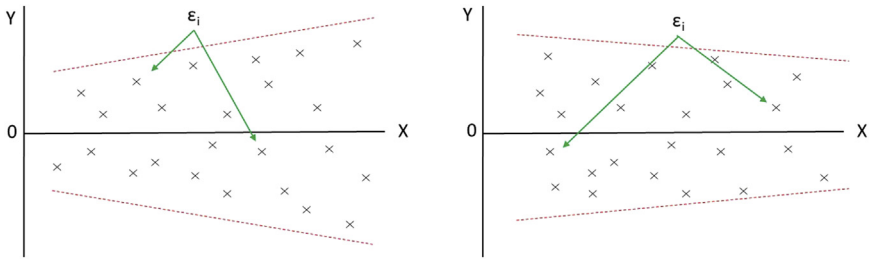

FIGURE 3.1 Overdispersed (left) and underdispersed (right) residuals.

Within the context of Poisson trials, there are numerous attributes that lead to overdispersion or unobserved heterogeneity (i.e., the crash rate that differs across observations). The crash process is very complex and involves numerous factors related to the interaction between the driver, the vehicle, and the environment (roadway and traffic characteristics), as discussed in Chapter 2—*Fundamentals and Data Collection*. Mannering et al. (2016) provide a list of factors that can lead to unobserved heterogeneity based on the interactions described earlier. They also discussed factors that influence crash severity, such as energy dissipation, impact angle, and the physiological characteristics of the vehicle occupants, once the crash occurred. The issue is that routinely collected data for safety analyses cannot capture all the factors that influence crash risk, even with the naturalistic data that are now being collected in the United States and Europe (see Chapter 12—*Data Mining and Machine Learning Techniques*). This will influence how much of the unobserved heterogeneity can be captured by statistical modeling. Later in the chapter, different models will be presented for minimizing overdispersion or reduce unobserved heterogeneity.

It is important to note that as the sample mean increases, and the number of zero observations becomessmaller, the characteristics of the crash data will tend toward equi-dispersion, irrespective of the missing factors that affect unobserved heterogeneity. This is particularly true for safety data that are analyzed at the county or state-level of analysis (Stevens and Lord, 2005; Blower et al., 2020). When the data are characterized by a very large mean, the model often will revert back to a "pure" Poisson process.

3.4.2 Underdispersion

In rare cases, crash data can exhibit underdispersion (Oh et al., 2006; Daniels et al., 2010). This can occur under two conditions. First, the data themselves exhibit underdispersion. This usually happens when the sample mean is very low and the data contain a large proportion of zero observations, which is also combined with a very small right-hand tail. With this characteristic, the data are almost characterized as a binomial distribution. The main attribute of the binomial distribution is that the sample variance is always smaller than the sample mean. The dataset by Oh et al. (2006) exhibits this characteristic, in which the mean-to-variance ratio is equal to 1.1. This dataset will be used for Exercise 3.3.

Second, the underdispersion is observed after the model is fitted with the data. In this case, the original data are often marginally overdispersed or may be equi-dispersed (such as datasets with a very large mean). At that point, when the model is fitted, that is, the observations are conditionally modeled upon the mean $(y|\mu)$, the modeling output shows the model's residuals exhibiting underdispersion (Lord et al., 2010). It is therefore important for the safety analyst to adequately explore the data for

identifying if there is the possibility that the modeling output could indicate underdispersion (see Chapter 5—*Exploratory Analyses of Safety Data*).

3.5 Basic count models

This section describes the basic models that have been proposed for analyzing safety data.

3.5.1 Poisson model

Because crash data are nonnegative integers, the application of standard ordinary least-squares regression (which assumes a continuous dependent variable) is not appropriate. Given that the dependent variable is a nonnegative integer, the Poisson regression model has initially been used as a starting point for analyzing crash data (Jovanis and Chang, 1986). In a Poisson regression model, the probability of a roadway entity (segment, intersection, etc.) i having y_i crashes per some time period (where y_i is a nonnegative integer) is given by

$$P(y_i|\mathbf{x}_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \quad \text{with} \quad y_i = 0, 1, 2, 3, \ldots, n \tag{3.5}$$

where $P(y_i|\mathbf{x}_i)$ is the probability of roadway entity (or observation) i having y_i crashes per unit of time and μ_i is the Poisson mean parameter for roadway entity i, which is equal to roadway entity i's expected number of crashes per year, $E[y_i]$. Poisson regression models are estimated by specifying the Poisson parameter μ_i (the expected number of crashes per period) as a function of explanatory variables, the most common functional form being $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, as described in Eq. (3.3).

Although the Poisson model has served as a starting point for crash-frequency analysis during the 80s, this model is rarely or no longer used for safety analyses for the reasons explained earlier. Furthermore, the Poisson model assumes that the crash risk (i.e., μ_i) is the same for all entities that have the same characteristics (same covariates), which is theoretically impossible.

3.5.2 Negative binomial model

The negative binomial (NB) (or Poisson-gamma) model is an extension of the Poisson model that was introduced to overcome the overdispersion observed in data. The NB model assumes that the Poisson parameter follows a gamma probability distribution. The model parameterization is characterized by a closed-form equation and the mathematics to manipulate the relationship between the mean and the variance structures is relatively simple. Because of these properties, the NB has become a very

popular model among practitioners and researchers working in numerous fields (Hilbe, 2011).

The NB model is derived by rewriting the Poisson parameter for each observation i with $E[y_i|\mathbf{x}_i] = \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i)$, where $\exp(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α or $\alpha = 1/\phi$. The addition of this term allows the variance to differ from the mean, such that $\mathrm{Var}[y_i|\mathbf{x}_i] = E[y_i|\mathbf{x}_i][1 + \alpha E[y_i|\mathbf{x}_i]] = E[y_i|\mathbf{x}_i] + \alpha E[y_i|\mathbf{x}_i]^2$. The Poisson model is a limiting model of the NB regression model as α approaches zero, which means that the selection between these two models is dependent upon the value of α. The parameter α is often called the dispersion parameter (or the over-dispersion parameter). The inverse of dispersion is called the "inverse dispersion parameter" and is denoted as ϕ. The dispersion parameter is usually assumed fixed, but can be made dependent on the covariates of the model (Miaou and Lord, 2003; Geedipally et al., 2009): $\alpha_i = \exp(\mathbf{z}_i'\boldsymbol{\gamma} + \varpi_i)$, where $\mathbf{Z}$ is a vector of covariates that may not necessarily be the same ones that are used for the mean value and $\boldsymbol{\gamma}$ is a $(p \times 1)$ vector of parameters. Chapter 6—*Cross-Sectional and Panel Studies in Safety* provides more details about the variance structures of crash-frequency models.

The probability mass function (PMF) of the negative binomial distribution has the following form:

$$P(y_i|\mathbf{x}_i, \alpha) = \frac{\Gamma(1/\alpha + y_i)}{\Gamma(1/\alpha) y_i!}\left(\frac{1/\alpha}{(1/\alpha) + \mu_i}\right)^{1/\alpha}\left(\frac{\mu_i}{(1/\alpha) + \mu_i}\right)^{y_i} \tag{3.6}$$

where as in Eq. (3.1), $P(y_i|\mathbf{x}_i, \alpha)$ is the probability of roadway (or other type of) entity i having y_i crashes per time period; Γ denotes the gamma distributed function; and all other variables are as previously defined. This parameterization is defined as NB-2 in the econometric literature (Cameron and Trivedi, 2013). Another parameterization called NB-1, which also exists but is very seldom used in the safety literature, is described in Appendix A.

Similar to other fields, the Poisson-gamma or NB model is the most frequently used model in crash-frequency modeling (Lord and Mannering, 2010). Although very popular, the model does have some limitations, most notably its inability to handle underdispersed data, and potential biases for the dispersion parameter when the data are characterized by the low sample mean values and small sample sizes, datasets with long tails and large percentage of zeros (see Lord, 2006; Lord et al., 2007).

3.5.3 Poisson-lognormal model

The Poisson-lognormal (PLN) model provides a similar parameterization as the negative binomial model, but the error term,$\exp(\varepsilon_i)$ is lognormal rather than gamma distributed. In this context, $\varepsilon_i \tilde{} \mathrm{Normal}(0, \sigma^2)$, which translates to $E[y_i|\mathbf{x}_i] = \mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \sigma^2/2 + \varepsilon_i)$. In terms of modeling, the

mean has to be adjusted to account for the added term associated with the variance. The variance of the estimated count is given as follows: $\mathrm{Var}(y_i|\mathbf{x}_i) = e^{\mu_i+\sigma^2/2} + \left[e^{\sigma^2} - 1\right]e^{2\mu_i+\sigma^2}$.

Compared to the Poisson-gamma model, the PLN model is more flexible than its counterpart, particularly for observations located at the tail end of the distribution, as seen in Fig. 3.2 (Khazraee et al., 2018). On the other hand, the NB tends to fit the data better near the zero counts. The PLN model is also more robust or reliable compared to the NB model

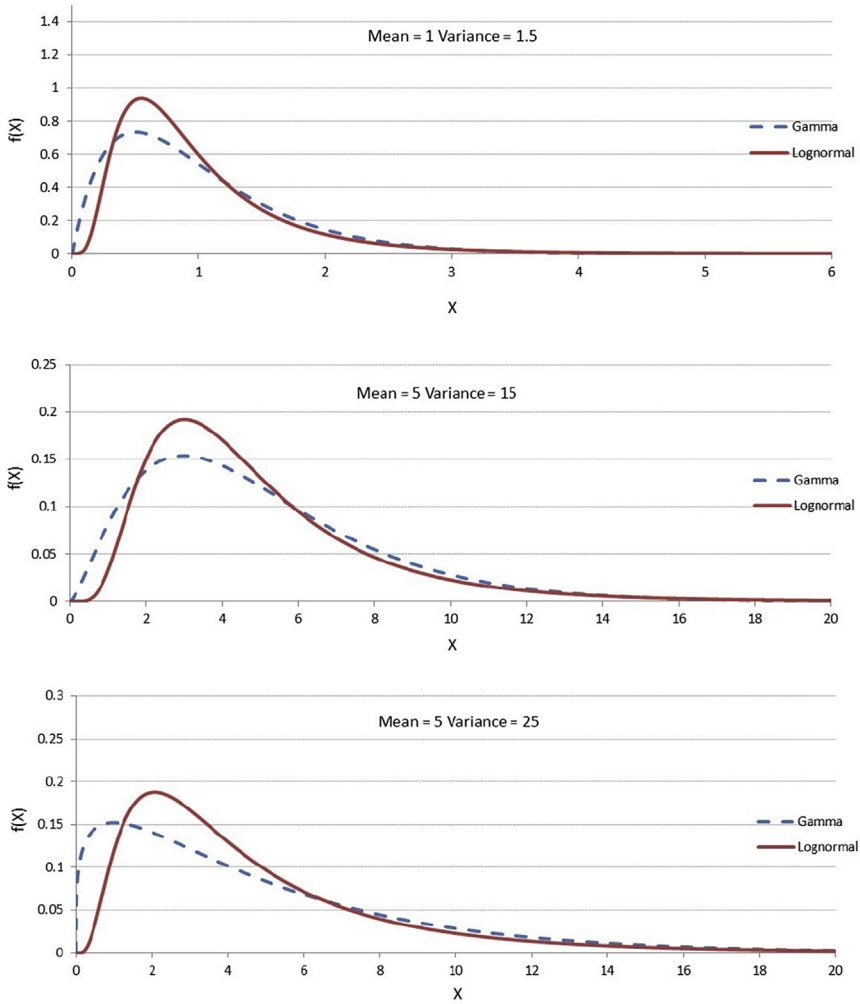

FIGURE 3.2 Probability distribution function of the Poisson-gamma (NB) and lognormal distributions for different mean-variance combinations (Khazraee et al., 2018).

for data characterized by low sample mean values and small sample size (Lord and Miranda-Moreno, 2008). With the latest advancements in Bayesian statistics, this model can be easily estimated or applied for analyzing safety data as the PMF of the model does not have a closed form. Codes already exist in WinBUGS and MATLAB for estimating Bayesian models (see Appendix C).

As discussed in the previous chapter, Shirazi and Lord (2019) described the boundaries when the PLN model should be selected over the NB model before fitting the data. The analyst needs to summarize the data for several types of statistics. They showed that the kurtosis and percentage-of-zeros in the data are among the most important summary statistics needed to distinguish between these two distributions. Fig. 3.3, reproduced from Chapter 2—*Fundamentals and Data Collection*, outlines the decision tree for selecting one model over the other. Exercise 3.1 provides an example about how to use this figure. Exercise 3.2 shows an example when the NB and PLN are compared using the same dataset.

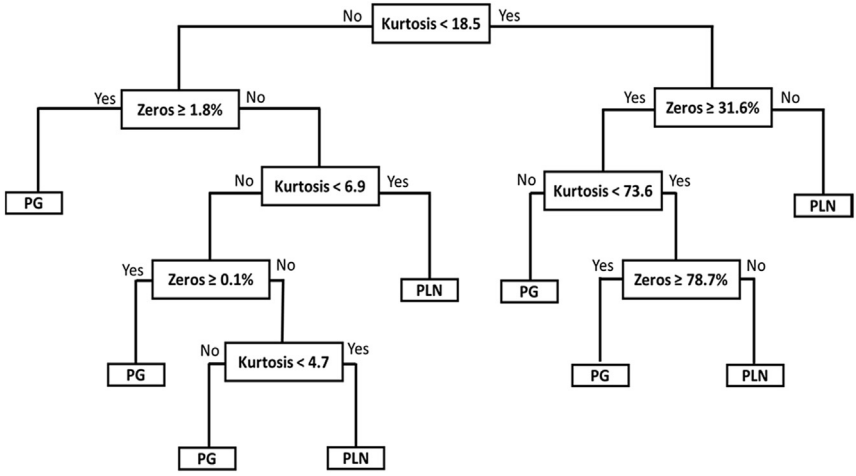

FIGURE 3.3 Characteristics-based heuristic to select a model between the PG and PLN distributions (Lord et al., 2019) (the tree can be used for data with the characteristics of $0.1 < \text{mean} < 20$ and $1 < \text{VMR} < 25$).

Exercise 3.1

Using the Texas Rural Two-Lane Highways Dataset, compare the characteristics of the data to determine which distribution between the NB and PLN appears to be more adequate.

First, using the information presented in Chapter 5, summarize the characteristics of this dataset. The summary statistics show that the percentage of zeros is 69.9% and the kurtosis is 123.6.

Second, examine the decision tree described in Fig. 3.3.

Looking at this tree, for a Kurtosis of 123.6 and the percentage of zeros equal to 69.9%, the selected model should be the PLN. The solution is found using the right-hand side of the tree.

Exercise 3.2

Estimate an NB model and a PLN model using the subset of the dataset used in Exercise 3.1 to compare the two models.

Use the following variables: Annual Average Daily Traffic (AADT) (veh/day), Lane width (LW) (ft), Speed limit (SL) (mph), and Left shoulder width (LSW) (ft). Use the segment length (L) (mile) as an offset, that is, the coefficient is equal to 1.

First, determine the functional form:

$$\mu = \exp\left(\beta_0 + \beta_1 \ln(F) + \sum_{j=2}^{p} \beta_j x_j\right)$$

The functional form can be manipulated to get the following:

$$\mu = \beta_0^* L F^{\beta_1} \exp\left(\sum_{j=2}^{p} \beta_j x_j\right)$$

The natural log of the AADT, $\ln(AADT)$, needs to be used to characterize the nonlinear relationship F^{β_1} and has to be taken out of the left-hand side of the equation. This form ensures that when the traffic flow is equal to zero, there is zero risk of a crash. The coefficient of the intercept is given as follows, $\beta_0^* = \exp(\beta_0)$.

continued

Exercise 3.2 (*cont'd*)

Second, as the PLN cannot be estimated using the maximum likelihood estimate (MLE), the coefficients are estimated using the Bayes method. The hyperparameters are estimated as follows:

Regression coefficients for both models: $\beta_j \sim \text{Uniform}(-\infty, +\infty)$, $j = 0, 1, 2, \ldots, p$;

Inverse dispersion parameter of PG: $\phi \sim \text{Uniform}(0, +\infty)$;

Dispersion parameter of PLN: $\sigma \sim \text{Uniform}(0, +\infty)$.

Third, estimate the posterior distribution of the coefficients using Markov Chain Monte Carlo (MCMC) simulation methods. The sampling from all parameters can be carried out using the Metropolis–Hastings algorithm or Slice Sampling (such as in WinBUGS software). Discard the first 20,000 samples, and estimate the posterior distribution of each coefficient using the next 60,000 MCMC iterations.

Fourth, present the results of the posterior estimates.

	Poisson-gamma			Poisson-lognormal		
Variable	**Estimate**	**Std. Error[a]**	**Pr(>|z|)**	**Estimate**	**Std. Error[a]**	**Pr(>|z|)**
Intercept (β_0)	−7.3166	1.6729	0.0000	−7.5750	1.6729	0.0000
Ln(AADT) (β_1)	0.5521	0.0949	0.0000	0.5477	0.082	0.0000
RSW (β_2)	0.0733	0.0695	0.2942	0.0710	0.755	0.9252
LW (β_3)	0.2366	0.1009	0.0184	0.2569	0.1068	0.0173
SL (β_4)	0.0193	0.0192	0.0201	0.0204	0.0201	0.3181
LSW (β_5)	−0.0285	0.0631	0.6513	−0.0246	0.0680	0.7202
ϕ	2.4660	1.1162	0.0284			
σ				0.6898	0.1001	0.0000
DIC	883.3			891.8		

[a]*Based on the credible intervals.*

The modeling results are almost similar, as the heuristic method showed that the PG is slightly favored over the PLN and the DIC is very close to each other. This was also confirmed with the heuristic method (the kurtosis was close to the boundary between the PG and PLN).

Exercise 3.2 (*cont'd*)

The final model for the Poisson-gamma is as follows:

$$\mu = (\exp(-7.3166))LF^{0.55}\exp(0.073RSW + 0.237LW + 0.019SL - 0.029LSW)$$

3.5.4 Other Poisson-mixture models

Following the introduction of the NB model, several other Poisson-mixture models have been examined or proposed in the context of crash data analysis. Some of these models include:

Poisson–Weibull: The Poisson–Weibull distribution model performs as well as the NB model and its coefficients can be easily estimated using the MLE (Connors et al., 2013; Cheng et al., 2013).

Poisson-Inverse Gaussian (PIG): The PIG model performs similar to the PLN, in that the model fits the data better at the tail end of the distribution (Zha et al., 2016). The coefficients are also easily estimated using the MLE.

Poisson-Inverse Gamma: This model also performs similar to the PLN for long-tailed data, as shown in Fig. 3.2 (Khazraee et al., 2018). The model is estimated using the Bayesian estimating method, which requires further work than the MLE.

Sichel [SI]: This model has been used or applied more frequently than the previous models (Wu et al., 2015). The SI model is recommended to be used for long-tailed data (Zou et al., 2015). This model can be estimated using the MLE.

Poisson–Tweedie: Depending on the parameterization of the model, the Poisson–Tweedie distribution models can become as special cases to the NB, PIG, or SI model (Debrabant et al., 2018). The same characteristics as those listed above apply here.

It should be noted that most of these models have not been evaluated for their stability when the safety data are characterized by the low sample mean and small sample biases (Lord, 2006).

3.6 Generalized count models for underdispersion

This section describes models that have been proposed for analyzing datasets that are either underdispersed or the modeling results exhibit underdispersion (conditioned upon the mean).

3.6.1 Conway–Maxwell–Poisson model

The most well-known model used for analyzing underdispersion is the Conway–Maxwell–Poisson (COM-Poisson). This model was originally proposed by Conway and Maxwell (1962) to analyze queues and service rates. Shmueli et al. (2005) further elucidated the statistical properties of the COM-Poisson distribution using the formulation given by Conway and Maxwell (1962), and Kadane et al. (2006) developed the conjugate distributions for the parameters of the COM-Poisson. The latter researchers proposed the following parameterization for the mean function of the model:

$$P(y_i|\mathbf{x}_i) = \frac{1}{Z(\lambda_i,\nu)}\frac{\lambda_i^{y_i}}{(y_i!)^{\nu}} \tag{3.7a}$$

$$Z(\lambda_i,\nu) = \sum_{n=0}^{\infty}\frac{\lambda^{\nu}}{(n!)^{\nu}} \tag{3.7b}$$

where y_i is the number of crashes per unit of time for observation i; λ_i is a centering parameter that is approximately equal to the mean of the observations in many cases (not exactly the same as μ) for observation i; and, ν is defined as the shape parameter of the COM-Poisson distribution. The COM-Poisson can model both underdispersed ($\nu > 0$) and overdispersed ($\nu < 0$) data, and several common PMFs are special cases of the COM with the original formulation. Specifically, setting $\nu = 0$ yields the geometric distribution; $\lambda < 1$ and $\nu \to \infty$ yields the Bernoulli distribution in the limit; and, $\nu = 1$ yields the Poisson distribution. The parametrization above leads to the following relationship for the mean and variance, respectively:

$$E[y_i|\mathbf{x}_i] = \mu_i \approx \lambda_i^{1/\nu + \frac{1}{2\nu} - \frac{1}{2}} \tag{3.8a}$$

$$\mathrm{Var}[y_i|\mathbf{x}_i] \approx \frac{1}{\nu}\lambda_i^{1/\nu}. \tag{3.8b}$$

The parameterization in Eq. (3.8) does not provide a clear centering parameter for the parameter λ. To overcome this issue, Guikema and Coffelt (2007) have proposed a reparameterization of the COM-Poisson distribution to provide a clear centering parameter. With the new parameterization, they were able to develop the generalized linear model (GLM) version of the distribution (see the steps in Lord et al., 2008). The equations are as follows:

$$P(y_i|\mathbf{x}_i) = \frac{1}{Z(\mu_i,\nu)}\left(\frac{\mu_i^{y_i}}{y_i!}\right)^{\nu} \tag{3.9a}$$

$$Z(\mu_i, \nu) = \sum_{n=0}^{\infty} \left(\frac{\mu_i^n}{n!} \right)^{\nu} \tag{3.9b}$$

$$\mu_i = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j \right) \tag{3.9c}$$

$$\nu_i = \exp\left(\gamma_0 + \sum_{l=1}^{q} \gamma_l x_l \right) \tag{3.9d}$$

where $\mu_i = \lambda_i^{1/\nu}$, the mean of the response variable. With this parameterization, the shape parameter can be made dependent on the covariates of the model. Similar to the NB with a varying dispersion, the covariates do not need to be the same as for the mean function.

Since its introduction, the COM-Poisson has become very popular not only in highway safety, but also in various areas, such as in management, economics, statistics, and ecology.

Exercise 3.3

Using the South Korean Dataset, estimate a COM-Poisson model.

Although old, this dataset collected in South Korea has been used for comparing models characterized with underdispersion. In this case, the underdispersion can be observed after the model is fitted, as the raw data showed near equal-dispersion (as discussed in the text earlier).

Use the following variables: AADT (veh/day), Average daily rail traffic (ADRT), Presence of commercial area (PCA), Train detector distance (mile) (TDD), Presence of track circuit controller (PTCC), Presence of guide (PG), and Presence of speed hump (PSH).

First, determine the functional form:

$$\mu = \beta_0^* F^{\beta_1} \exp\left(\sum_{j=2}^{p} \beta_j x_j \right)$$

In this functional form, F is the AADT flow on the segment. The natural log of the AADT, $\ln(AADT)$, needs to be used to characterize the nonlinear relationship F^{β_1}.

Second, estimate the coefficient of the model using the MLE (see attached codes[a]).

continued

Exercise 3.3 (*cont'd*)

Third, present the results of the model.

Variable	Estimate	Std. error	Pr(>|z|)
Intercept (β_0)	−7.075	1.286	0.0000
Ln(AADT) (β_1)	0.658	0.141	0.0000
ADRT (β_2)	0.005	0.004	0.2112
PCA (β_3)	1.500	0.515	0.0041
TDD (β_4)	0.002	0.001	0.0456
PTCC (β_5)	−1.214	0.437	0.0063
PG (β_6)	−0.994	0.513	0.0522
PSH (β_7)	−1.571	0.539	0.0048
ν	2.365		
Log likelihood	−94.88		

[a] *https://cran.r-project.org/web/packages/COMPoissonReg/COMPoissonReg.pdf*.

The modeling results show that the model output is heavily underdispersed ($\nu = 2.365$).

3.6.2 Other generalized models

Other models that have been proposed for analyzing underdispersion include the following:

Gamma model (continuous distribution): This model cannot account for observations with zero counts. It is presented here as a caution for not using this model.

Gamma-count model: This modified gamma model has been proposed by Winkelman (1995). The parameterization offered by this researcher assumes that the observations have a direct correlation with each other in time. In safety, this means that a crash at time t is directly related to a crash at time $t + n$, which is again theoretically impossible.

Double Poisson: This model has initially been proposed by Efron (1986). However, it has not been used often, as the normalizing constant of the model is not properly definite. Zou et al. (2013) have proposed a different parameterization of the constant term and found results similar to the COM-Poisson.

Hyper-Poisson: The hyper-Poisson (hP) is a two-parameter generalization of the Poisson distribution. Similar to the COM-Poisson, it can model the variance function as a function of the covariates. It performs as well as the COM-Poisson and can be estimated using the MLE (Khazraee et al., 2015).

Generalized Event Count: This model uses the theoretical statistics called "bilinear recurrence relationship" that was introduced by Katz (1965) for describing the dispersion parameter of the Poisson count model. Ye et al. (2018) applied the model to crash data and found its performance to be similar to the hP.

3.7 Finite mixture and multivariate models

This section describes models that have been proposed to examine unknown subpopulations within datasets (finite mixture models) and when different dataset attributes are correlated and need to be accounted for when analyzed simultaneously (multivariate models).

3.7.1 Finite mixture models

Finite mixture models are a type of models that can be utilized to examine heterogeneous populations. For these models, the assumption states that the overall data are generated from several distributions that are mixed together, with the underlying principle that individual observations are generated from an unknown number of distributions or subpopulations. Similar to the NB model, the Finite Mixture Negative Binomial or FMNB is also the most popular model among finite mixture models (note: the notation FM is placed in front of the Poisson or NB model designation).

There are several reasons to expect the existence of different subpopulations as the crash data are generally collected from various geographic, environmental, and geometric design contexts over some fixed time period. In such cases, it may be inappropriate to apply one aggregate NB model, which could lead to the misinterpretation of the modeling results. Therefore, it is reasonable to hypothesize that individual crashes occurring on highways or other safety-related entities are generated from a certain number (K) of hidden subgroups, or components that are unknown to the transportation safety analyst. For the FMNB model, the final outputs include the number of components, component proportions, component-specific regression coefficients, and the degree of overdispersion within each component.

For the FMNB-K model, it is assumed that the marginal distribution of y_i follows a mixture of negative binomial distributions (Park and Lord, 2009),

$$P(y_i|\mathbf{x}_i,\boldsymbol{\Theta}) = \sum_{k=1}^{K} w_k \mathrm{NB}(\mu_{k,i},\phi_k) = \sum_{k=1}^{K} w_k \left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\Gamma(\phi_k)}\left(\frac{\mu_{k,i}}{\mu_{k,i}+\phi_k}\right)^{y_i}\left(\frac{\phi_k}{\mu_{k,i}+\phi_k}\right)^{\phi_k}\right]. \tag{3.10}$$

With the expected value and variance given by

$$E(y_i|\mathbf{x}_i,\boldsymbol{\Theta}) = \sum_{k=1}^{K} w_k \mu_{k,i} \tag{3.11}$$

$$\mathrm{Var}(y_i|\mathbf{x}_i,\boldsymbol{\Theta}) = E(y_i|\mathbf{x}_i,\boldsymbol{\Theta}) + \left(\sum_{k=1}^{K} w_k \mu_{k,i}^2 \left(1+{}^{1}/_{\phi_k}\right) - E(y_i|\mathbf{x}_i,\boldsymbol{\Theta})^2\right) \tag{3.12}$$

where $\mu_{k,\ i} = \exp(\mathbf{x}_i\boldsymbol{\beta}_k)$ for subgroup; $\boldsymbol{\beta} = (\beta_1,\beta_2,\ldots,\beta_K)'$ denotes the vector of all parameters; and, $\mathbf{w} = (w_1,w_2,\ldots,w_K)'$ refers to a weight distribution of which elements are restricted to be positive and sum to unity ($w_k > 0$ and $\sum w_k = 1$). In this case, even if all the component's means are the same, the variance of y_i will always be greater than the mean.

This type of model has become increasingly popular among highway safety analysts as this model can be used for identifying potential sources of dispersion by way of mixed distributions. It should be noted that zero-inflated models are a special case of finite mixture models, where one of the two latent classes has a long-term mean equal to zero, which is theoretically impossible (Park and Lord, 2007). Zero-inflated models are briefly discussed in the last section of this chapter.

3.7.2 Multivariate models

Multivariate models are used when different crash severities or collision types are analyzed simultaneously. These models are needed as the severity levels and collision types may not be independent. As shown by Kuo and Lord (2020), the location of fatal crashes for example may also be associated with crashes that are not fatal, but lead to incapacitated injuries (note: the crashes themselves are not correlated). In this case, the factors, such as the speed limit, may explain why these severity types are related geographically. Treating the correlated crash counts as independent and applying a univariate model to each category can lead to less precise estimates for estimating the effects of factors on crash risk. Fortunately, multivariate models can account for the correlation between crash counts directly into the modeling process.

This type of model type has the same basic nomenclature as the univariate model. The difference is that the vectors are in fact matrices $(p \times m)$, where p refers to the number of parameters in the model and m corresponds to the number of different crash severity or collision types. The notation for multivariate models is as follows (Park and Lord, 2007):

$$\mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nm} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_{11} & \cdots & y_{1p} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pm} \end{bmatrix}. \tag{3.13}$$

With the matrices above, the model is defined (always assuming the crash count is Poisson distributed) (Park and Lord, 2007):

$$P(y_{im}|\mathbf{x}_i, b_i, \beta_m) \tilde{} \text{Poisson}(\mu_{im}) \tag{3.14}$$

where

$$\mu_{im} = \exp(x_i \beta_m + b_{im}) \tag{3.15}$$

for $m = 1, \ldots, M$ and $i = 1, \ldots, n$. The y_{im}'s are assumed independent given the μ_{im}'s. To model the correlations among the crash counts of M different severity or collision types, let

$$P(b_i|\Sigma) \tilde{} N_M(0, \Sigma) \tag{3.16}$$

where $\sum$ is an unrestricted covariance matrix; and, N_M denotes M-dimensional multivariate normal distribution. It was shown in Chib and Winkelmann (2001) that the variance of y_{im} is greater than the mean (allowing for overdispersion) as long as the diagonal elements of $\sum$ are greater than 0, and the covariance between the counts, y_{im} and ν_{im} can be positive or negative depending on the sign of the mth element of $\sum$. Thus, the correlation structure of the crash counts is unrestricted.

In the safety literature, the multivariate model has been proposed for both the Poisson-gamma (MVNB) (Ma and Kockleman, 2006) and the Poisson-lognormal (MVPLN) (Park and Lord, 2007) models. However, the correlation matrix for the MVNB model cannot be negative, which provides less flexibility than for the MVPLN. Hence, since their introduction, the MVPLN has been the only multivariate model used for analyzing safety data.

3.8 Multi-distribution models

Multi-distribution models are different from multivariate models as they include multiple shape/scale parameters or mix more than two distributions simultaneously (e.g., Poisson-gamma-Bernoulli). On

occasions, some of these models have been called as zero-centered models as the majority of the distribution weight lies near zero. Technically, multi-distribution models could also be manipulated to become multivariate models, but so far, no researchers have developed the combination of these two categories of models.

3.8.1 Negative Binomial–Lindley model

The Negative Binomial–Lindley (NB-L), as the name implies, is a mixture of the NB and the Lindley distributions (Lindley, 1958; Ghitany et al., 2008). This three-parameter distribution has interesting properties in which the distribution is characterized by a single long-term mean that is never equal to zero and a single variance function, similar to the traditional NB distribution (note: as discussed below, the third parameter is used to refine the estimate of the gamma parameter of the NB model).

Before tackling the NB-L, recall that NB distribution can be parameterized in two different manners—either as a mixture of the Poisson and gamma distributions (described earlier) or based on a sequence of independent Bernoulli trials. Using the latter parameterization, the PMF of the NB distribution can be given as follows:

$$P(y_i|\phi,q)=\frac{\Gamma(\phi+y_i)}{\Gamma(\phi)\times y_i!}(q)^{\phi}(1-q)^{y_i};\phi>0,\ \ 0<q<1 \tag{3.17}$$

In this formulation, "q" is defined as the probability of failure in each trial, given by

$$q=\frac{\phi}{\mu_i+\phi} \tag{3.18}$$

where μ_i and ϕ, are as defined earlier. We can conveniently denote the above distribution as $NB(y_i|\phi,\mu_i)$. We can now define the NB-L distribution as a mixture of the NB and Lindley distributions as follows:

$$P(y_i|\mu_i,\phi,\theta)=\int \mathrm{NB}(y_i|\phi,\varepsilon_i\mu_i)\ \ \mathrm{Lindley}(\varepsilon_i|\theta)d\varepsilon_i \tag{3.19}$$

Notice that, μ_i, the mean of the NB distribution is multiplied by a random term ε_i, which follows the Lindley distribution. The Lindley distribution is given as:

$$f(x|\theta)=\frac{\theta^2}{\theta+1}(1+x)e^{-\theta x};\theta>0,\ \ x>0 \tag{3.20}$$

where θ is the shape parameter.

Under the assumption that the number of crashes y follows an NB-L (ϕ, q) distribution, then the mean function is given as follows:

$$E(y_i) = \mu_i \times E(\varepsilon_i) = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) \frac{\theta+2}{\theta(\theta+1)} \tag{3.21}$$

The variance can be obtained as follows:

$$\mathrm{Var}(y_i) = \mu_i \times \frac{\theta+2}{\theta(\theta+1)} + \mu_i^2 \times \frac{2(\theta+3)}{\theta^2(\theta+1)} \times \frac{(1+\phi)}{\phi} - \left(\mu_i \times \frac{\theta+2}{\theta(\theta+1)}\right)^2 \tag{3.22}$$

Notice that the structural form of the regression is comparable to that of in Eq. (3.3), if we rearrange the terms as $\beta_j' = \beta_j + \log\left(\frac{\theta+2}{\theta(\theta+1)}\right)$.

For the GLM, the NB-L distribution, conditional upon the unobserved site-specific frailty term ε_i that explains additional heterogeneity, can be rewritten as follows (Geedipally et al., 2012):

$$\begin{aligned} P(y_i, \mu_i, \phi|\varepsilon_i) &= \mathrm{NB}(y_i|\phi, \varepsilon_i\mu_i) \\ \varepsilon_i &\tilde{} \mathrm{Lindley}(\varepsilon_i|\theta). \end{aligned} \tag{3.23}$$

The above formulation can be thought of as an instance of the Generalized Linear Mixed model where the mixed effects follow the Lindley distribution. However, considering that the Lindley is not a standard distribution, the hierarchical representation of the Lindley distribution can be further utilized.

As defined by Zamani and Ismail (2010), the Lindley distribution is a two-component mixture given by

$$\varepsilon \tilde{} \frac{1}{1+\theta}\mathrm{Gamma}(2,\theta) + \frac{1}{1+\theta}\mathrm{Gamma}(1,\theta). \tag{3.24}$$

Recognizing the special structure in the mixture components, the above equation can be rewritten as follows:

$$\varepsilon_i \tilde{} \sum \mathrm{Gamma}(1+z_i,\theta)\mathrm{Bernoulli}\left(z_i \middle| \frac{1}{1+\theta}\right). \tag{3.25}$$

Under a Bayesian framework, a hierarchical representation of the Lindley distribution can be represented as

$$\varepsilon_i \tilde{} \mathrm{Gamma}(1+z_i,\theta) \tag{3.26a}$$

$$z_i \sim \text{Bernoulli}\left(\frac{1}{1+\theta}\right) \tag{3.26b}$$

whose marginal distribution is the Lindley distribution. The complete multilevel hierarchical model can now be given as follows (Geedipally et al., 2012):

$$P(y_i, \mu_i, \phi|\varepsilon_i) = \text{NB}(y_i|\phi, \varepsilon_i\mu_i) \tag{3.27a}$$

$$\varepsilon_i \sim \text{Gamma}(\varepsilon_i|1+z_i, \theta) \tag{3.27b}$$

$$z_i \sim \text{Bernoulli}\left(z_i\middle|\frac{1}{1+\theta}\right) \tag{3.27c}$$

The hierarchical model mixes three distributions: Poisson, gamma, and Bernoulli.

Since its introduction, the NB-L model has been successfully applied for analyzing numerous datasets (see Appendix B for a list of studies). In all cases, the NB-L worked very well to capture excess zero responses. Recently, a random-parameters NB-L (RPNB-L) model was introduced for analyzing safety data, which further helps reducing unobserved heterogeneity (random-parameters models are described later) (Shaon et al., 2018).

Shirazi et al. (2017) indicated, using the heuristics method, that the NB-L is usually preferred over the NB when the skewness of the data is larger than 1.92. As the boundary was calculated via heuristics methods, the difference in selecting one distribution or model over another will become more significant if the value is further away from the boundary (say 2.5 or 3).

Exercise 3.4

Using the Texas Rural Divided Multilane Highways Dataset, compare the characteristics of the data to determine which distribution between the NB and NB-L appears to be more adequate.

First, using the methods described in Chapter 5, summarize the characteristics of the Texas data and find the skewness. The skewness is equal to 4.0.

Second, as the skewness is larger than 1.92, the NB-L is preferred over the NB. This is confirmed when other goodness-of-fit measures are used to compare the distributions, as shown in the following table (Sharazi et al., 2017).

Exercise 3.4 (*cont'd*)

Method	NB	NB-L	Criteria	Favored Distribution
Chi-square (χ^2)	2.73	1.68	$\chi^2_{NB-L} < \chi^2_{NB}$	NB-L
Log-likelihood (LL)	−696.1	−695.1	$LL_{NB-L} > LL_{NB}$	NB-L
DT heuristic	–		Skewness>1.92	NB-L
RF heuristic	–		Using the RF heuristic tool	NB-L

3.8.2 Other multi-distribution models

Other models that are categorized as multivariate models include the following:

Negative Binomial-Generalized Exponential (Vangala et al., 2015): This model is very similar to the NB-L, but the Lindley distribution is replaced by the generalized exponential (GE) distribution. The NB-GE is defined as follows:

$$P(y_i|\mu_i,\phi,\pi) = \int \mathrm{NB}(y_i|\phi,\varepsilon_i\mu_i)\ \mathrm{GE}(z_i|\pi,\tau)d\varepsilon_i \tag{3.28}$$

The GE distribution is given by (Aryuyuen and Bodhisuwan, 2013) as follows:

$$f(\varepsilon_i|\pi,\tau) = \pi\tau(1-e^{\tau\varepsilon_i})^{\pi-1}e^{-\tau\varepsilon};\ \pi,\tau>0,\ \ \varepsilon_i>0 \tag{3.29}$$

where π is the shape parameter, and τ is the scaled parameter, respectively.

This model has been found to perform very similar to the NB-L (Vangala et al., 2015) both in terms of goodness-of-fit and computing effort.

The Negative Binomial–Crack Model (Saengthong and Bodhisuwan, 2013): The negative binomial-Crack (NB-CR) distribution is obtained by

mixing the NB distribution with a CR distribution. The NB-CR distribution can be defined such that

$$P(y_i|\mu_i,\phi,\alpha) = \int \mathrm{NB}(y_i|\phi,\varepsilon_i\mu_i)\ \ \mathrm{CR}(z_i|\lambda,\theta,\gamma)d\varepsilon_i. \tag{3.30}$$

The parameter μ_i is the mean response for the number of crashes and is assumed to have a log-linear relationship with the covariates. The parameter ε_i follows the CR distribution. This model has not been evaluated using crash data.

3.9 Models for better capturing unobserved heterogeneity

This section describes models that have been proposed to better capture and understand unobserved heterogeneity. The two principle models are the random-effects and random-parameter models.

3.9.1 Random-effects/multilevel model

In the models described in Section 3.5, it is assumed that the size of the effect of the variables is fixed, meaning that one true effect size exists for all the observations and the difference, that is the unobserved heterogeneity between observations is a purely random error. In practice, the unobserved heterogeneity may not be solely attributed to a purely random error, but could also be partly explained by the differences in the observations themselves and between groups or levels of observations (as described in Section 3.4). These are called within observations variance (pure random error) and between observations variance, respectively. In highway safety, for example, this could mean that sites located within the same geographical areas are expected to share similar characteristics (say more homogeneous population within a neighborhood), as opposed to sites located in other geographical neighborhoods within a very large city, and this needs to be accounted for in the modeling effort.

Random-effects (RE) models, or sometimes called multilevel models (Gelman and Hill, 2007), are models that allow the variance that may exist within different levels of the data to be better depicted. This is accomplished by adding one or more RE terms or random intercept term to capture the between observations variance. Taking the basic models described in Section 3.5, the formulation becomes

$$\mu_{io} = \exp\left(\mathbf{x}'_{io}\boldsymbol{\beta} + \varpi_o + \varepsilon_{io}\right) \tag{3.31}$$

where μ_{io} is the mean of the observation i belonging to group o and ϖ_o is a random-effect or intercept term for group o. The end result of the random-

effects model is that it modifies the mean of the observation i by changing the value of the intercept. There are different parameterizations of the random-effects model and the random effect term can also be used to account for temporal correlation, such as for panel data (covered in Chapter 6—*Cross-Sectional and Panel Studies in Safety*), or analyzing safety data accross clusters in the recently introduced cross-classified random effects modeling (CCREM) (see (Bakhshi and Ahmed, 2021).

3.9.2 Random-parameters model

Random-parameters (RP) models can be viewed as an extension of random-effects models. However, rather than only influencing the intercept of the model, random-parameters models allow each estimated parameter of the model to vary across each individual observation in the dataset (after they are identified as being random). These models attempt to account for the unobserved heterogeneity from one roadway site to another (Milton et al., 2008). Over the last decade, this type of model has become quite popular.

These models, as crash-frequency models, have two general types of parameterization: (1) random parameters and (2) random parameters with means as a function of explanatory variables. A third parameterization that incorporates the potential correlation between the variables exists, but is not covered here, as it has not been frequently used in practice, with a few exceptions (see Saeed et al., 2019; Matsuo et al., 2020). They are described as follows (Mannering et al., 2016; (Washington et al., 2020)).

3.9.2.1 Random parameters

The RP model follows the same characteristics as for the fixed model, that is, $\mu_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i\right)$, but in this case, the coefficients are allowed to vary from one observation to the next. Thus, the coefficient k for each observation i in the vector $\mathbf{x}_i$ can be defined as

$$\beta_{ik} = b_k + \upsilon_{ik} \tag{3.32}$$

where β_{ik} is the coefficient for the explanatory variable k and observation i, b_k is the mean fixed parameter estimate across all observations for the explanatory variable k, and υ_{ik} is the randomly distributed variable that is used to capture unobserved heterogeneity across all observations. This term can assume any specified distribution, such as the normal or the gamma distribution.

3.9.2.2 Random parameters with means as a function of explanatory variables

The parameterization described in Eq. (3.32), which assumes a single mean $\beta_{ik} = b_k + v_{ik}$ across all observations, can be expanded by making the coefficient dependent on the explanatory variables themselves. In other words, the mean may directly vary from observation to observation as of the function of the observation's characteristics. In this regards, Eq. (3.32) can be rewritten as follows (Mannering et al., 2016):

$$\beta_i = \mathbf{b} + \mathbf{\Theta}\mathbf{z}_i \tag{3.33}$$

where $\mathbf{z}_i$ is a $(k \times 1)$-vector of explanatory variables for observation i that influence the random parameter vector, $\mathbf{\Theta}$ is a matrix $(k \times k)$ of estimable coefficients. As explained by Mannering et al. (2016, p. 9), "each row of $\mathbf{\Theta}$ corresponds to the loadings of a specific element of the β_i vector on the $\mathbf{z}_i$ vector; if a specific column entry in a row of $\mathbf{\Theta}$ is zero, it implies that there is no shift in the mean of the corresponding row element of the β_i vector due to the row element of the $\mathbf{z}_i$ vector corresponding to the column under consideration." This parameterization could help seize additional heterogeneity that a model with fixed coefficients could not. For example, drivers of different age groups could be driving vehicles that have unique characteristics, such as the age of the vehicle, including or not including specific safety features, which are not commonly used by other groups. In this regard, it is well known that younger drivers often drive older vehicles (hand-me downs from parents)[2].

RP models are powerful models that have become very popular in highway safety, in major part because they fit the data better (as they significantly reduce the unobserved heterogeneity). Although the models tend to provide a better fit (compared to fixed-parameter models), the coefficients could be difficult to interpret. For instance, one may have difficulties explaining why the variable "lane width" for a given value is associated with an increase in the crash risk for some sites, while the majority of the other sites are not. Furthermore, recent work by Tang et al., 2019 and Huo et al., 2020 showed that RP models may not be adequate for predicting crashes (say at new locations), at least with the datasets used in these two studies. Nonetheless, more work should be conducted on this topic.

[2] https://www.irishtimes.com/life-and-style/motors/older-smaller-cars-putting-younger-drivers-at-risk-1.2060380: "A US-based report has shown that teenagers and younger drivers are dying because their cars are older and less safe…. …The ages of these cars are observed by the report to be as much as 11-years, while 82% of teenage drivers killed in this period were driving a car of at least 6 year old. The researchers behind the report flagged up the fact that older, smaller cars were clearly less safe than newer, larger vehicles." (accessed December 2020).

It is instructive to note that the distinction between RE and RP models seems to be more pronounced in traffic safety modeling. Even though both modeling approaches focus on different structural aspects to reduce unobserved heterogeneity, both can be seen as special cases of a more general class of models called generalized linear latent and mixed models (GLLAMs) (McCulloch et al., 2008; Stroup, 2013; Skrondal et al., 2004). Consequently, much of the inference machinery that has been the mainstay in fitting GLLAMs can be used here. However, with the introduction of either RE or RP, the choice of the inference technique is not straightforward. Sometimes it can get philosophical as well. One can estimate the parameters using either the Frequentist method or the Bayesian method. One can also focus on marginal estimates or conditional estimates. Under the frequentist method, the RP models can include anywhere from a single parameter to all the parameters that will be identified as random (although extremely rare). As the likelihood function does not have a closed form in this case, simulation needs to be used for estimating the coefficients. Green (2004) used a Monte Carlo (MC) simulation, not MCMC, for estimating the parameters.

On the other hand, under the Bayesian estimation method, all the parameters are always defined as random by the nature of the Bayesian theory (Yamrubboon et al., 2019). A variety of Bayesian inference techniques can be employed, with the MCMC simulation technique being the most commonly used by statisticians. Occasionally, some RP models may be equivalent to a Bayesian hierarchical model.

In terms of reducing the unobserved heterogeneity, it should be pointed out that RP models reduce the heterogeneity via the parameters themselves, while multi-distribution models reduce it by decreasing the modeling error (via multiple shape parameters). If the ultimate (and misguided) objective is to solely reduce unobserved heterogeneity, then multi-distribution models are usually better than RP models (Shaon et al., 2018). However, as discussed at the end of the chapter, solely using this objective is not recommended.

3.10 Semi- and nonparametric models

This section describes semi- and nonparametric models. Semiparametric models are somewhat similar to parametric models in a sense that the crash counts are assumed to be Poisson distributed, but the Poisson mean and/or the coefficients of the model are assumed to follow a nonparametric distribution (referred to as distribution-free). Nonparametric models are models that are completely distribution free.

3.10.1 Semiparametric models

As described earlier, semiparametric models assume that the crash counts are Poisson distributed. There are different categories, but the two most common categories that have gained popularity in highway safety consist of estimating the Poisson mean or the coefficients of the model using smooth functions (e.g., spline, etc.). We can find two examples in the following.

The first one is known as the generalized additive models (GAMs). For this type of model, each coefficient is characterized by a distinct smooth function. Using the characteristics of Eq. (3.34), the relationship between the mean and the parameters can be defined as follows (Xie and Zhang, 2008):

$$\mu_i = \exp\left(\beta_0 + \sum_{j=1}^{p} f_j(x_{ij})\right) \tag{3.34}$$

where β_0 is the intercept of the model and $f_j(x_{ij})$ is the smooth function (e.g., P-splines, kernel cubic regression splines, smoothers, and thin-plate regression splines). The generalized additive models can also include a combination of fixed, nonlinear functions or a combination of nonlinear functions:

$$\mu_i = \exp\left(\beta_0 + \sum_{j=1}^{k} x_{ij} + f_{k12}\left(x_{i(k+1)}, x_{i(k+2)}\right) + \sum_{j=k+3}^{p} f_j(x_{ij})\right) \tag{3.35}$$

where $\sum_{j=1}^{k} x_{ij}$ is the parametric component of the model, $f_{k12}\left(x_{i(k+1)}, x_{i(k+2)}\right)$ is a smooth function taking two input variables, and $\sum_{j=k+3}^{p} f_j(x_{ij})$ is the summation of nonparametric smooth functions.

Fig. 3.4 shows an example of the application of the GAM model using crash and traffic flow data collected at signalized intersections in Toronto, Ontario. This dataset has been used extensively in the safety literature (Lord, 2000; Miaou and Lord, 2003). The functional form of the GAM is as follows:

$$\mu_i = \exp(\beta_0 + f_1(x_{i1}) + f_2(x_{i2})) \tag{3.36}$$

where $f_1(x_{i1})$ and $f_2(x_{i2})$ are the entering flows for the major and minor approaches, respectively. The relationship between the number of crashes and flow is illustrated in Fig. 3.4.

The second example is known as the seminonparametric (SNP) Poisson model (Ye at al., 2018). For this model, the unobserved heterogeneity captured by the model's error (ε) follows a K-not polynomial function.

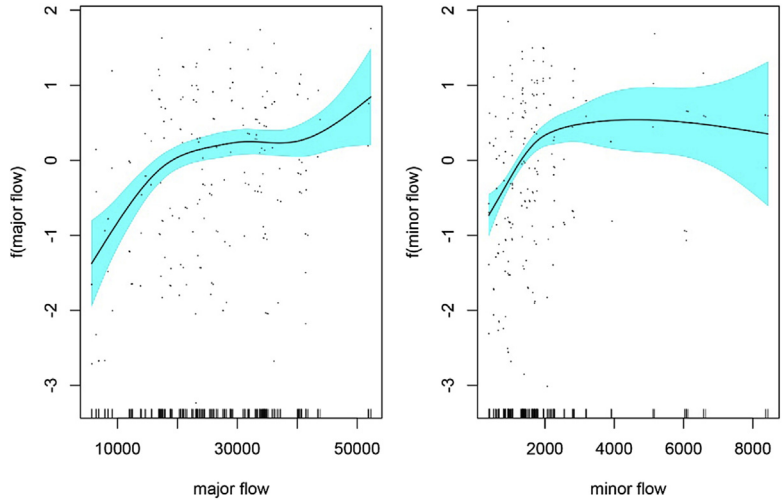

FIGURE 3.4 Relationship between the number of crashes and entering flows (Xie and Zhang, 2008).

The polynomial, called the SNP distribution, can be used to approximate different distributions, such as the normal, gamma, or binomial distribution among others. The probability density function (PDF) of the error term is given as follows (Ye et al., 2018):

$$f(\varepsilon) = \frac{\left(\sum_{m=0}^{k} a_m \varepsilon^m\right)^2 \phi(\varepsilon)}{\int_{-\infty}^{+\infty} \left(\sum_{m=0}^{k} a_m \varepsilon^m\right)^2 \phi(\varepsilon) d\varepsilon} \tag{3.37}$$

where "K" refers to the length of the polynomial, "m" is an indicator increasing from 0 to "K", a_m is a constant coefficient, and $\phi(\varepsilon)$ represents the PDF of the standard normal distribution. The function $\int_{-\infty}^{+\infty} f(\varepsilon) d(\varepsilon)$ in the denominator guarantees that it is equal to 1.

The denominator of the PDF described in Eq. (3.37) can be expanded by including another indicator "n" that also increases from 0 to K (Ye et al., 2018):

$$\int_{-\infty}^{+\infty} \left(\sum_{m=0}^{k} a_m \varepsilon^m\right)^2 \phi(\varepsilon) d\varepsilon = \sum_{m=0}^{K} \sum_{n=0}^{K} a_m a_n \int_{-\infty}^{+\infty} \varepsilon^{m+n} \phi(\varepsilon) d\varepsilon \tag{3.38}$$

Using the recursive equation $I(n) = \int_{-\infty}^{+\infty} \varepsilon^n \phi(\varepsilon) d\varepsilon$, where $I(0) = 1$, $I(1) = 0$, and $I(n) = n - 1 \times I(n-2)$, when $n \geq 2$, the denominator of Eq. (3.38) can be defined such that

$$\int_{-\infty}^{+\infty} \left(\sum_{m=0}^{k} a_m \varepsilon^m\right)^2 \phi(\varepsilon) d\varepsilon = \sum_{m=0}^{K} \sum_{n=0}^{K} a_m a_n I(m+n) \tag{3.39}$$

By mixing the SNP with the Poisson distribution, the SNP-Poisson model can be defined as follows:

$$P(y_i|\mathbf{x}_i) = \int_{-\infty}^{+\infty} P(y_i|\varepsilon_i) f(\varepsilon_i) d\varepsilon_i$$

$$P(y_i|\mathbf{x}_i) = \int_{-\infty}^{+\infty} \left\{ \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i} \times \frac{\left(\sum_{m=0}^{k} a_m \varepsilon_i^m\right)^2 \phi(\varepsilon_i)}{\sum_{m=0}^{K}\sum_{n=0}^{K} a_m a_n I(m+n)} \right\}. \tag{3.40}$$

As the unconditional probability function of Eq. (3.40) does not have a closed form, the numerical method of the Gauss–Hermite quadrature needs to be applied to approximate the unconditional probability:

$$P(y_i|\mathbf{x}_i) = \sum_{j=1}^{J} \left\{ w_j \left\{ \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i} \right\} \times \left[\frac{\left(\sum_{m=0}^{k} a_m \varepsilon_i^m\right)^2 \phi(\varepsilon_i)}{\sum_{m=0}^{K}\sum_{n=0}^{K} a_m a_n I(m+n)} \right] \right\}. \tag{3.41}$$

As discussed in Press et al. (2007), the Gaussian quadrature is a cutting-edge procedure that can correctly evaluate the integrals in the likelihood function with a small number of supporting points. Ye et al. (2018) presented a table listing the weights and the number of nodes needed to characterize a Poisson-gamma model. They have also showed that, depending on the number of nodes, the model can reliably replicate the Poisson-gamma and Poisson-lognormal models. Another advantage of this model is that it can capture multimodal datasets, similar to the finite mixture models previously described.

3.10.2 Dirichlet process models

Models that employ the Dirichlet process (DP), widely used in the Bayesian literature, can technically be classified as either nonparametric or semiparametric models depending on the modeling framework (Antoniak, 1974; Escobar and West, 1995). For semiparametric models, as applied with safety data, the count data still follow a Poisson distribution, but the mean or the error term is assumed to follow a Dirichlet distribution or process. As opposed to the Poisson or NB mixtures, in which two parametric distributions are mixed together, the DP is characterized by an infinite mixture of distributions, where the number of unique components or distributions and the component characteristics themselves can be learned from the data (Shirazi et al., 2016).

The DP (Ferguson, 1973, 1974) is a stochastic process that is usually used as a prior in Bayesian nonparametric (or semiparametric) modeling. In this regard, Escobar and West (1998) defined the DP as a random probability measure over the space of all probability measures. In that

sense, the DP is considered as a distribution over all possible distributions; that is, each draw from the DP is itself a distribution, which may not be same as for the previous draw.

As described in Shirazi et al. (2016), let $A_2, \ldots, A_r$ $A_1, A_2, \ldots, A_r$ be any finite measurable partitions of the parameter space (Θ). Let us assume τ be a positive real number and $F_0(.|\theta)$ be a continuous distribution over Θ. Then, $F() \sim \mathrm{DP}(\tau, F_0(.|\theta))$ if and only if (Escobar and West, 1998):

$$(F(A_1), F(A_2), \ldots, F(A_r)) \sim \mathrm{DP}(\tau F_0(A_1|\theta), \tau F_0(A_2|\theta), \ldots, \tau F_0(A_r|\theta)) \quad (3.42)$$

where τ is defined as the precision (or concentration) parameter and $F_0(.|\theta)$ as the base (or baseline) distribution. Note that, based on the Dirichlet distribution properties, each partition $A \subset \Theta$ is defined as follows:

$$E(F(A)) = F_0(A|\theta) \quad (3.43)$$

$$\mathrm{Var}(F(A)) = \frac{F_0(A|\theta)(1 - F_0(A|\theta))}{1 + \tau}. \quad (3.44)$$

Therefore, the base distribution $F_0(.|\theta)$ and the precision parameter τ play significant roles in the DP definition. The expectation of the random distribution $F()$ is the base distribution $F_0(.|\theta)$. Likewise, the precision parameter τ controls the variance of the random distribution around its mean. In other words, τ measures the variability of the target distribution around the base distribution. As $\tau \to \infty$, $F() \to F_0(.|\theta)$ while, on the other hand, as $\tau \to 0$, the random distribution $F()$ would deviate further away from $F_0(.|\theta)$.

Eq. (3.45) defines the DP indirectly through the marginal probabilities assigned to a finite number of partitions. Therefore, it gives no intuition on realizations of $F() \sim \mathrm{DP}(\tau, F_0(.|\theta))$. To simulate random distributions from the DP, however, Sethuraman (1994) introduced a straightforward stick-breaking constructive representation of this process as follows:

$$\gamma_k|\tau \sim \mathrm{Beta}(1, \tau), \quad k = 1, 2, \ldots \quad (3.45a)$$

$$\psi_k|\theta \sim F_0(.|\theta), \quad k = 1, 2, \ldots \quad (3.45b)$$

$$p_k = \gamma_k \prod_{k'<k} (1 - \gamma_{k'}), \quad k = 1, 2, \ldots \quad (3.45c)$$

$$F() \sim \mathrm{DP}(\tau, F_0(.|\theta)) \equiv \sum_k p_k \delta_{\psi_k} \quad (3.45d)$$

where δ_{ψ_k} indicates a degenerate distribution with all its mass at ψ_k. This construction, metaphorically, can be considered as breaking a unit length of stick iteratively (Ishwaran and James, 2001). First, the stick is broken at a random proportion γ_1; an atom is generated from the base distribution (ψ_1) and is assigned to the length of the stick that was just broken (p_1). Then, recursively, the remaining portions of the stick are broken at new proportions $(\gamma_2, \gamma_3, \ldots)$; new atoms are generated from the base

distribution $(\psi_2, \psi_3, \ldots)$ and are assigned to each broken length of the remained sticks $(p_2, p_3, \ldots)$.

Given the stick-breaking construction of the DP (Eq. 3.45), the mean and variance of $\upsilon \tilde{} F()$ can be calculated as follows (Yang et al., 2010):

$$E(\nu|p,\psi) = \mu_{\mathrm{DP}} = \sum_k p_k \psi_k \tag{3.46}$$

$$\mathrm{Var}(\nu|p,\psi) = \nu_{\mathrm{DP}} = \sum_k p_k \psi_k^2 - \left(\sum_k p_k \psi_k\right)^2. \tag{3.47}$$

As indicated in Eq. (3.45), theoretically, the stick-breaking construction of the DP includes infinite components (so called clusters); however, practically, the model can be approximated with its truncated version (TDP) by considering an upper bound on the number of components (M) as follows (Ishwaran and James, 2001; Ishwaran and Zarepour, 2002):

$$\gamma_k|\tau \tilde{} \mathrm{Beta}(1,\tau), \quad k = 1,2,\ldots,M \tag{3.48a}$$

$$\psi_k|\theta \tilde{} F_0(.|\theta), \quad k = 1,2,\ldots,M \tag{3.48b}$$

$$p_k = \gamma_k \prod_{k'<k} (1-\gamma_{k'}), \quad k = 1,2,\ldots,M \tag{3.48c}$$

$$F() \tilde{} TDP(\tau, F_0(.|\theta)) \equiv \sum_k p_k \delta_{\psi_k}. \tag{3.48d}$$

So far, several research studies have tried to estimate the required number of components (or clusters) (M) in the truncated version of the DP (Ishwaran and James, 2001; Ohlssen et al., 2007). As a key point, first, the analyst needs to keep in mind that the number of mass points (M) in the TDP is correlated to the value of the precision parameter (τ). Theoretically, as the value of τ increases, the number of clusters that are shared by data points increases; hence, a larger value for the parameter M is required. Second, the model needs to be approximated to the level that it can be assumed that the effect of neglected clusters remains negligible $\left(1 - \sum_{k=1}^{M} p_k \approx \varepsilon\right)$. Given these two rationales into account, Ohlssen et al. (2007) showed that the maximum number of clusters can be approximated by Eq. (3.49) as a function of τ and the desired ε- accuracy as follows:

$$M \approx 1 + \frac{\log(\varepsilon)}{\log\left(\frac{\tau}{1+\tau}\right)} \tag{3.49}$$

Once the model is approximated to M clusters, p_M needs to be modified using Eq. (3.50) to make the model identifiable (i.e., $\sum_{k=1}^{M} p_k = 1$):

$$p_M = 1 - \sum_{k=1}^{M-1} p_k \tag{3.50}$$

In the safety literature, two semiparametric models have been proposed, one for the Poisson (Heydari et al., 2016), called the Poisson–Dirichlet Process (P-DP), and one for the multi-parameter NB model (Shirazi et al., 2016), called the NB-DP. For the P-DP, we have this formulation:

$$P(y|\mu,\tau,F(.|\theta)) = \int \text{Poisson}(y|\nu\mu)dF(\nu|\text{DP}(\tau,F(.|\theta))). \tag{3.51}$$

For the NB-DP, we have the following:

$$P(y|\mu,\phi,\tau,F(.|\theta)) = \int \text{NB}(y|\nu\mu,\phi)dF(\nu|\text{DP}(\tau,F(.|\theta))). \tag{3.52}$$

As stated in the previous section, the distribution of $\text{DP}(\tau,F(.|\theta))$ can be approximated by its truncated construction $\text{TDP}(\tau,M,F(.|\theta))$. Consequently, the P-TDP and NB-TDP model can be seen as a hierarchical Bayesian model described as follows:

$$P(y_i|\nu_i\mu_i) = \text{Poisson}(\nu_i\mu_i) \tag{3.53a}$$

$$P(y_i|\nu_i\mu_i,\phi_i) = \text{NB}(\nu_i\mu_i,\phi) \tag{3.53b}$$

$$\gamma_k|\tau \sim \text{Beta}(1,\tau), \quad k=1,2,\ldots,M \tag{3.53c}$$

$$\psi_k|\theta \sim F_0(.|\theta), \quad k=1,2,\ldots,M \tag{3.53d}$$

$$p_k = \gamma_k \prod_{k'<k}(1-\gamma_{k'}), \quad k=1,2,\ldots,M \tag{3.53e}$$

$$\nu_i \sim F() \tag{3.53f}$$

$$F() \sim \text{TDP}(\tau,F_0(.|\theta)) \equiv \sum_k p_k \delta_{\psi_k} \tag{3.53g}$$

where $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, again as described earlier.

The NB-DP or NB-TDP has been shown to be very flexible with datasets characterized with a large amount of zero observations and/or long tail. Shirazi et al. (2017) evaluated the model using a dataset from Indiana, which has a long tail. Table 3.1 shows the comparison results with the NB and NB-L.

A by-product of the NB-TDP model, data can be classified into a finite number of clusters. This clustering property is based on how different sites share the mixed effect mass points (v). To benefit from the advantage

TABLE 3.1 Modeling Results for the Indiana Data (Shirazi et al., 2017).

Variables	NB			NB-L			NB-TDP		
	Estimate	**Std. Error**	**Pr(>|z|)**	**Estimate**	**Std. error**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**Pr(>|z|)**
Intercept (β_0)	−4.779	0.979	0.0000	−3.739	1.115	0.0008	−7.547	0.1227	0.0000
Ln(ADT) (β_1)	0.722	0.091	0.0000	0.630	0.106	0.0000	0.983	0.117	0.0000
Friction (β_2)	−0.02774	0.008	0.0006	−0.02746	0.011	0.1300	−0.01999	0.008	0.0126
Pavement (β_3)	0.4613	0.135	0.0005	0.4327	0.217	0.0468	0.3942	0.152	0.0100
MW (β_4)	−0.00497	0.001	0.0000	−0.00616	0.002	0.0021	−0.00468	0.002	0.0195
Barrier (β_5)	−3.195	0.234	0.0000	−3.238	0.326	0.0000	−8.035	1.225	0.0000
Rumble (β_6)	−0.4047	0.131	0.0021	−0.3976	0.213	0.0609	−0.3780	0.150	0.0134
$\alpha = 1/\phi$	0.934	0.118	0.0000	0.238	0.083	0.0074	0.301	0.085	0.0042
DIC[a]		1900			1701			1638	
MAD[b]		6.91			6.89			6.63	
MSPE[c]		206.79			195.54			194.5	

[a] *Deviance information criterion.*
[b] *Mean absolute deviance.*
[c] *Mean squared predictive error.*

Site	1	2	3	4	5	6	7	8	9	10
1	1.0	0.6	0.6	0.6	0.6	0.2	0.6	0.6	0.1	0.1
2	0.6	1.0	0.6	0.6	0.6	0.2	0.6	0.6	0.1	0.1
3	0.6	0.6	1.0	0.6	0.6	0.2	0.6	0.6	0.1	0.1
4	0.6	0.6	0.6	1.0	0.6	0.2	0.6	0.6	0.1	0.1
5	0.6	0.6	0.6	0.6	1.0	0.2	0.6	0.6	0.1	0.1
6	0.2	0.2	0.2	0.2	0.2	1.0	0.2	0.2	0.6	0.6
7	0.6	0.6	0.6	0.6	0.6	0.2	1.0	0.6	0.1	0.1
8	0.6	0.6	0.6	0.6	0.6	0.2	0.6	1.0	0.1	0.1
9	0.1	0.1	0.1	0.1	0.1	0.6	0.1	0.1	1.0	0.6
10	0.1	0.1	0.1	0.1	0.1	0.6	0.1	0.1	0.6	1.0

FIGURE 3.5 The heatmap representation of the partitioning matrix for the top 10 sites with the highest ADT values in the Indiana dataset (Shirazi et al., 2018).

of clustering, the partitioning information matrix needs to be recorded at each iteration of the MCMC. The matrix can be used to investigate similarities between sites especially with regards to recognizing unobserved variables or identifying safety issues and deploying countermeasures. An example is presented in Fig. 3.5.

Fig. 3.5 shows the heatmap representation of the partitioning matrix for the top 10 sites with the highest ADT values. The figure shows the likelihood that site "X" and "Y" fall into the same cluster. For simplicity, the probabilities were rounded to the first decimal. A higher likelihood will be represented by a darker shade on the map. As observed in this figure, for instance, with relatively high probability (~60%), site "1" falls into the same cluster as site "2", site "3" or several more. This information can offer insights to identify potential unobserved variables or safety issues and decide on appropriate countermeasures for the site "1."

3.10.3 Nonparametric models

Nonparametric models have been used relatively often in highway safety. Some recent popular models include the multilayer perceptron

(MLP) neural network (Xie et al., 2007; Kononov et al., 2011), convolutional neural networks (Ma et al., 2017), Bayesian neural networks (BNN) (Xie et al., 2007), and support vector machine (SVM) (Li et al., 2008). These models and other models are described in greater detail in Chapter 12—*Data Mining and Machine Learning Techniques*, which discusses methods related to the analysis of naturalistic and other types of data.

For crash-count analyses, MLP, BNN, and SVM have so far been used for predicting crashes rather than for examining relationships (Singh et al., 2018; Dong et al., 2019). They are very good at predicting crashes, but they, unfortunately, work as black boxes. Hence, the safety analyst needs to be very familiar with the characteristics of the data and the assumptions associated with their use. Because of the nature of these models (i.e., nonparametric), they could easily over-fit the data. Techniques, such as sensitivity analyses, can be used for assessing the relationship between the independent variables and the dependent variable (crash count). For example, Fish and Blodgett (2003) have proposed a method for examining the sensitivity of MLP models. Their method can also be used for assessing the variables of BNN and SVMs (Xie et al., 2007; Li et al., 2008).

3.11 Model selection

In this chapter, we have presented different types of models that vary from the most basic to very complex. Positive and negative attributes have been provided for most of these models, along with a description explaining when the model could be suitable given the known and unknown characteristics of the data. In the highway safety literature, a lot of work has been devoted to the development and application of statistical models (in fact, this is the majority of the work produced in highway safety according to Zou and Vu, 2019). A common theme of a new study is that the "new" proposed model is claimed to be better than other previously published or widely applied models because it fits the data better. In other words, the "new" model reduces the unobserved heterogeneity more than the previous model. Although this could be a legitimate objective, the final selection of the model, as discussed earlier, should not solely be based on how good the model fits the data. The model also needs to adequately capture the data generating process of the dataset under study. Miaou and Lord (2003) refer to this subject as the "goodness-of-logic," which consists of making sure the model properly characterizes the analyzed data and the model is methodologically sound. For example, the safety literature has shown that zero-inflated (ZI) and Hurdle NB models usually provide a better statistical fit than the traditional NB. However, the main assumption for this model is that one of the two states' key attributes shows that the long-term mean is equal to zero (i.e., the Poisson

mean is equal to zero) (note: for the ZI model, a proportion of the 0s come from a Poisson mean equal to zero; whereas, for the Hurdle model, all the 0s come from such distribution). It is obviously not possible to observe sites that could never experience any crashes despite road users traveling on the facilities[3] (Lord et al., 2005, 2007). This fundamental and other related issues have also been raised in environmental science (Warton, 2005), social sciences (Allison, 2012), substance abuse (Xie et al., 2013), and criminology (Fisher et al., 2017; Britt et al., 2018). Solely looking at the fit could allow the safety analyst to miss important cues or information, such as coefficients that are counterintuitive or hidden correlation among variables.

Along the same line, using a very complex model does not necessarily mean that the model is better, even if the "fit" is superior. The model could be overly complex given the study objectives or the gains they provide compared to traditional models considered marginal. As Dr. George Box famously said "all models are wrong but some are useful" (Box, 1979) and "the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity" (Box, 1976), meaning that more complex models are not necessarily better. Furthermore, complex models often do not have the opportunity to be fully validated using theoretical principles, simulation, or a wide range of datasets, especially as many have been recently been introduced in the safety literature. It is not uncommon to see new models that are later found to suffer from important methodological limitations (e.g., zero-inflated models). Depending on the parameterization and the estimation method, the model could also take a very long time to provide results. For example, some models estimated using the Bayesian method can sometimes take several days or hours for the MCMC posterior estimates to converge. A pragmatic or better approach would be to use the MLE method when it can be used based on the study objectives and characteristics of the data, but to use the Bayesian

[3] Lambert (1992), who first introduced zero-inflated models, clearly stated that these models are used for two distinct categories of observations: perfect state and nonperfect state. The idea behind this kind of model is to assign a probability for these observations with 0s to fall on either one of these two states. She also mentioned that these models can be difficult to interpret. This limitation is even addressed in the abstract. Note that the use of the Vuong statistic (Vuong, 1989) to evaluate when these models should be used has also been criticized by other researchers (e.g., Wilson, 2015). Pew et al. (2020) argued that zero-inflated models could still be used for predicting crashes, irrespective if the relationship between the number of crashes and risk factors makes sense or not. More details about the inadequacy of zero-inflated models in different fields can be found on statistician and sociologist Dr. Paul D. Allison's website: https://statisticalhorizons.com/zero-inflated-models?

estimation method when the MLE cannot be used because of the complexity of the model.

Although the NB model can become unreliable under particular conditions, this model, which is characterized by solid theoretical foundations, has had the opportunity to be analyzed, tweaked and used by researchers and practitioners across the globe for several decades and is considered more than adequate for most applications (Hilbe, 2014). Hence, safety analysts should not automatically reject this basic model for analyzing crash data. However, if the study objectives, as described in Chapter 2—*Fundamentals and Data Collection*, and the characteristics of the data are such that a better model may be more suitable, then an alternative model should be selected (where the "fit" plays an important role, but is not the sole decision factor). The same principles apply to the crash-severity models described in the next chapter.

References

Allison, P., 2012. Logistic Regression Using SAS: Theory and Application, second ed. SAS Institute Inc., Cary, NC. See Do we really need zero-inflated models? Statistical Horizons (Blog by Emeritus Sociology Professor Dr. Paul Allison) Retrieved from http://statisticalhorizons.com/zero-inflated-models. (Accessed 29 July 2020).

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Stat. 2 (6), 1152–1174.

Aryuyuen, S., Bodhisuwan, W., 2013. Negative binomial-generalized exponential (NB-GE) distribution. Appl. Math. Sci. 7 (22), 1093–1105.

Bakhshi, A.K., Ahmed, M.M., 2021. Practical advantage of crossed random intercepts under Bayesian hierarchical modeling to tackle unobserved heterogeneity in clustering critical versus non-critical crashes. Accid. Anal. Prev. 149, 105855.

Blower, D., Flannagan, C., Geedipally, S., Lord, D., Wunderlich, R., 2020. Identification of Factors Contributing to the Decline of Traffic Fatalities in the United States from 2008 to 2012, vol. 928. NCHRP RESEARCH REPORT, Washington, D.C.

Britt, C.L., Rocque, M., Zimmerman, G.M., 2018. The analysis of bounded count data in criminology. J. Quant. Criminol. 34, 591–607. https://link.springer.com/article/10.1007/s10940-017-9346-9?shared-article-renderer.

Box, G.E.P., 1976. Science and statistics. J. Am. Stat. Assoc. 71 (356), 791–799. https://doi.org/10.1080/01621459.1976.10480949.

Box, G.E.P., 1979. Robustness in the strategy of scientific model building. In: Launer, R.L., Wilkinson, G.N. (Eds.), Robustness in Statistics. Academic Press, pp. 201–236. https://doi.org/10.1016/B978-0-12-438150-6.50018-2, 9781483263366.

Cameron, A.C., K Trivedi, P., 2013. Regression Analysis of Count Data, second ed. Cambridge University Press, N.Y.

Cheng, L., Geedipally, S.R., Lord, D., 2013. Examining the Poisson-Weibull generalized linear model for analyzing crash data. Saf. Sci. 54, 38–42.

Chib, S., Winkelmann, R., 2001. Markov chain monte carlo analysis of correlated count data. J. Bus. Econ. Stat. 19, 428–435.

Connors, R.D., Maher, M., Wood, A., Mountain, L., Ropkins, K., 2013. Methodology for fitting and updating predictive accident models with trend. Accid. Anal. Prev. 56, 82–94.

Conway, R.W., Maxwell, W.L., 1962. A queuing model with state dependent service rates. J. Ind. Eng. Int. 12, 132–136.

Daniels, S., Brijs, T., Nuyts, E., Wets, G., 2010. Explaining variation in safety performance of roundabouts. Accid. Anal. Prev. 42 (2), 393–402.
Davis, G.A., 2000. Accident reduction factors and causal inference in traffic safety studies: a review. Accid. Anal. Prev. 32 (1), 95–109.
Debrabant, B., Halekoh, U., Bonat, W.H., Hansen, D.L., Hjelmborg, J., Lauritsen, J., 2018. Identifying traffic accident black spots with Poisson-Tweedie models. Accid. Anal. Prev. 111, 147–154.
Dong, C., Xie, K., Sun, X., Lyu, M., Yue, H., 2019. Roadway traffic crash prediction using a statespace model based support vector regression approach. PLoS One 14 (4) e0214866.
Efron, B., 1986. Double exponential families and their use in generalized linear Regression. J. Am. Stat. Assoc. 81 (395), 709–721.
Elvik, R., 2003. Assessing the validity of road safety evaluation studies by analysing causal chains. Accid. Anal. Prev. 35 (5), 741–748.
Elvik, R., 2011. Assessing causality in multivariate accident models. Accid. Anal. Prev. 43, 253–264.
Escobar, M.D., West, M., 1995. Bayesian density-estimation and inference using mixtures. J. Am. Stat. Assoc. 90 (430), 577–588.
Escobar, M.D., West, M., 1998. Computing nonparametric hierarchical models. Pract. Nonparametric Semiparametric Bayesian Stat. 1–22.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Ann. Stat. 209–230.
Ferguson, T.S., 1974. Prior distributions on spaces of probability measures. Ann. Stat. 615–629.
Fish, K.E., Blodgett, J.G., 2003. A visual method for determining variable importance in an artificial neural network model: an empirical benchmark study. J. Target Meas. Anal. Market. 11 (3), 244–254.
Fisher, W.H., Hartwell, S.W., Deng, X., 2017. Managing inflation: on the use and potential misuse of zero-inflated count regression models. Crime Delinquen. 63 (1), 77–87.
Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-lindley generalized linear model: characteristics and application using crash data. Accid. Anal. Prev. 45 (2), 258–265.
Geedipally, S.R., Lord, D., Park, B.-J., 2009. Analyzing different parameterizations of the varying dispersion parameter as a function of segment length. Trans Res Rec 2103 108–118.
Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge, UK.
Ghitany, M.E., Atieh, B., Nadarajah, S., 2008. Lindley distribution and its application. Math. Comput. Simul. 78, 39–49.
Green, W., 2004. Interpreting Estimated Parameters and Measuring Individual Heterogeneity in Random Coefficient Models. Working paper. Department of Economics, Stern School of Business, New York University.
Guikema, S.D., Coffelt, J.P., 2007. A flexible count data regression model for risk analysis. Risk Anal. 28 (1), 213–223.
Hauer, E., 2010. Cause, effect and regression in road safety: a case study. Accid. Anal. Prev. 42, 1128–1135.
Hauer, E., 2020. Crash causation and prevention. Accid. Anal. Prev. 143, 105528.
Heydari, S., Fu, L., Lord, D., Mallick, B.K., 2016. Multilevel Dirichlet process mixture analysis of railway grade crossing crash data. Anal. Methods Accid. Res. 9, 27–43.
Hilbe, J.M., 2011. Negative Binomial Regression, second ed. Cambridge University Press, Cambridge, UK.
Hilbe, J.M., 2014. Modeling Count Data. Cambridge University Press, Cambridge, UK.

Huo, X., Leng, J., Hou, Q., Zheng, L., Zhao, L., 2020. Assessing the explanatory and predictive performance of a random parameters count model with heterogeneity in means and variances. Accid. Anal. Prev. 147, 105759. https://doi.org/10.1016/j.aap.2020.105759.
Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. 96 (453), 161–173.
Ishwaran, H., Zarepour, M., 2002. Exact and approximate representations for the sum Dirichlet process. Can. J. Stat. 30 (2), 269–283.
Jovanis, P.P., Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled. Transp. Res. Rec. 1068, 42–51.
Kadane, J.B., Shmueli, G., Minka, T.P., Borle, S., Boatwright, P., 2006. Conjugate analysis of the Conway-Maxwell-Poisson distribution. Bayesian Anal 1, 363–374.
Katz, L., 1965. Unified treatment of a broad class of discrete probability distributions. Class. Contag. Discret. Distrib. 1, 175–182.
Khazraee, S.H., Johnson, V., Lord, D., 2018. Bayesian Poisson hierarchical models for crash data analysis: investigating the impact of model choice on site-specific predictions. Accid. Anal. Prev. 117, 181–195.
Khazraee, S.H., Saez-Castillo, A.J., Geedipally, S.R., Lord, D., 2015. Application of the hyper-Poisson generalized linear model for analyzing motor vehicle crashes. Risk Anal. 35 (5), 919–930.
Kononov, J., Lyon, C., Allery, B., 2011. Relation of flow, speed, and density of urban freeways to functional form of a safety performance function. Transp. Res. Rec. J. Transp. Res. Board (2236), 11–19.
Kuo, P.-F., Lord, D., 2020. Applying the colocation quotient index to crash severity analyses. Accid. Anal. Prev. 135, 105368.
Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34 (1), 1–14.
Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accid. Anal. Prev. 40 (4), 1611–1618.
Lindley, D.V., 1958. Fiducial distributions and Bayes' theorem. J. R. Stat. Soc. Series B Stat. Methodol. 20 (1), 102–107.
Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accid. Anal. Prev. 38 (4), 751–766.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. A 44 (5), 291–305.
Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Saf. Sci. 46 (5), 751–770.
Lord, D., Guikema, S.D., Geedipally, S., 2008. Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accid. Anal. Prev. 40 (3), 1123–1134.
Lord, D., Geedipally, S.R., Guo, F., Jahangiri, A., Shirazi, M., Mao, H., Deng, X., 2019. Analyzing Highway Safety Datasets: Simplifying Statistical Analyses from Sparse to Big Data. Report No. 01-001, Safe-D UTC. U.S. Department of Transportation, Washington, D.C.
Lord, D., Persaud, B.N., 2000. Accident Prediction Models with and without Trend: Application of the Generalized Estimating Equations (GEE) Procedure, vol. 1717. Transportation Research Record, pp. 102–108.
Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37 (1), 35–46.

Lord, D., Washington, S.P., Ivan, J.N., 2007. Further notes on the application of zero inflated models in highway safety. Accid. Anal. Prev. 39 (1), 53–57.
Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., Wang, Y., 2017. Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. Sensors 17 (4), 818.
Ma, J., Kockelman, K.M., 2006. Bayesian multivariate Poisson regression for models of injury count, by severity. In: Transportation Research Record: Journal of the Transportation Research Board. No. 1950, pp. 24–34.
Mannering, F., Shankar, V., Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.
Matsuo, K., Sugihara, M., Yamazaki, M., Mimura, Y., Yang, J., Kanno, K., Sugiki, N., 2020. Hierarchical Bayesian modeling to evaluate the impacts of intelligent speed adaptation considering individuals' usual speeding tendencies: a correlated random parameters approach. Anal. Methods Accid. Res. 27, 100125.
McCulloch, C.E., Searle, S.R., Neuhaus, J.M., 2008. Generalized, Linear, and Mixed Models, second ed. John Wiley & Sons Inc., Hoboken, NJ.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall/CRC, Boca Raton, FL.
Miaou, S.-P., Lord, D., 2003. Modeling traffic-flow relationships at signalized intersections: dispersion parameter, functional form and Bayes vs empirical Bayes. Transp. Res. Rec. 1840, 31–40.
Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accid. Anal. Prev. 40 (1), 260–266.
Molnar, C., 2020. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable eBook. https://christophm.github.io/interpretable-ml-book/index.html.
Myers, R., 2000. Classical and Modern Regression with Applications, second ed. Duxbury Press, Belmont, U.S.
Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway–highway interfaces. Accid. Anal. Prev. 38 (2), 346–356.
Ohlssen, D.I., Sharples, L.D., Spiegelhalter, D.J., 2007. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. Stat. Med. 26 (9), 2088–2112.
Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41 (4), 683–691.
Park, E.S., Lord, D., 2007. Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. Transportation Research Record, pp. 1–6.
Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 2007. Numerical recipes. In: The Art of Scientific Computing, third ed. Cambridge University Press.
Pearl, J., 2009. Casualty: Causality Models, Reasoning and Inference, second ed. Cambridge University Press, Cambridge, UK.
Saeed, T.U., Hall, T., Baroud, H., Volovski, M.J., 2019. Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: an empirical assessment of multilane highways. Anal. Methods Accid. Res. 23, 100101.
Saengthong, P., Bodhisuwan, W., 2013. Negative binomial-crack (NB-CR) distribution. J. Pure Appl. Math. 84, 213–230.
Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Stat. Sin. 4 (2), 639–650.
Shaon, M.R.R., Qin, X., Shirazi, M., Lord, D., Geedipally, S., 2018. Developing a random parameters negative binomial-lindley model to analyze highly over-dispersed crash count data. Anal. Methods Accid. Res. 18, 33–44.

Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. A methodology to design heuristics for model selection based on characteristics of data: application to investigate when the negative binomial lindley (NB-L) is preferred over the negative binomial (NB). Accid. Anal. Prev. 107, 186–194.

Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over dispersed count data with a heavy tail: characteristics and applications to crash data. Accid. Anal. Prev. 91, 10–18.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P., 2005. A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. J. R. Stat. Soc. Ser. C. 54, 127–142.

Singh, G., Sachdeva, S.N., Pal, M., 2018. Support vector machine model for prediction of accidents on non-urban sections of highways. Proc. Inst. Civ. Eng. Transp. 171 (5), 253–263.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, FL.

Stroup, W.W., 2013. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Chapman & Hall/CRC, Boca Raton, FL.

Tang, H., Gayah, V.V., Donnell, E.T., 2019. Evaluating the predictive power of an SPF for two-lane rural roads with random parameters on out-of-sample observations. Accid. Anal. Prev. 132, 105275.

Vangala, P., Lord, D., Geedipally, S.R., 2015. Exploring the application of the negative binomial-generalized exponential model for analyzing traffic crash data with excess zeros. Anal. Methods Accid. Res. 7, 29–36.

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57, 307–333.

Warton, D.I., 2005. Many zeros does not mean zero inflation: Comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16 (2), 275–289.

Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. Statistical and Econometric Methods for Transportation Data Analysis, third ed. CRC Press - Taylor & Francis Group, New York, N.Y.

Wilson, P., 2015. The misuse of the Vuong test for non-nested models to test for zero-inflation. Econ. Lett. 127, 51–53.

Winkelmann, R., 1995. Duration dependence and dispersion in count-data models. J. Bus. Econ. Stat. 13 (4), 467–474.

Wu, L., Lord, D., Zou, Y., 2015. Validation of CMFs derived from cross sectional studies using regression models. Transp. Res. Rec. 2514, 88–96.

Xie, Y., Lord, D, Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural networks: An empirical analysis. Accid. Anal. Prev. 39 (5), 922–933.

Xie, H., Tao, J., McHugo, G., Drake, R.E., 2013. Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: an example of smoking cessation. J. Subst. Abuse Treat. 45, 99–108.

Xie, Y., Zhang, Y., 2008. Crash frequency analysis with generalized additive models. Transp. Res. Rec. J. Transp. Res. Board 39–45. No. 2061.

Yamrubboon, D., Thongteeraparp, A., Bodhisuwan, W., Jampachaisri, K., Volodin, A., 2019. Bayesian inference for the negative binomial-sushila linear model. Lobachevskii J. Math. 40 (1), 42–54.

Yang, M.A., Dunson, D.B., Baird, D., 2010. Semiparametric Bayes hierarchical models with mean and variance constraints. Comput. Stat. Data Anal. 54 (9), 2172–2186.

Ye, X., Wang, K., Zou, Y., Lord, D., 2018. A semi-nonparametric Poisson regression model for analyzing motor vehicle crash data. PLoS One 13 (5) e0197338.

Ye, Z., Xu, Y., Lord, D., 2018. Crash data modeling with a generalized estimator. Accid. Anal. Prev. 117, 340–345.

Zamani, H., Ismail, N., 2010. Negative binomial-Lindley distribution and its application. J. Math. Stat. 6 (1), 4–9.
Zha, L., Lord, D., Zou, Y., 2016. The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data. J. Transp. Saf. Secur. 8 (1), 18–35.
Zou, Y., Geedipally, S.R., Lord, D., 2013. Evaluating the double Poisson generalized linear model. Accid. Anal. Prev. 59, 497–505.
Zou, X., Vu, H.L., 2019. Mapping the knowledge domain of road safety studies: a scientometric analysis. Accid. Anal. Prev. 132, 1, 05243.
Zou, Y., Wu, L., Lord, D., 2015. Modeling over-dispersed crash data with a long tail: examining the accuracy of the dispersion parameter in negative binomial models. Anal. Methods Accid. Res. 5–6, 1–16.