CHAPTER

6

Cross-sectional and panel studies in safety

6.1 Introduction

Transportation agencies spend numerous resources to better understand crash contributing factors and to determine appropriate safety countermeasures to reduce traffic crashes and fatalities. Observational and experimental studies are the two study designs that are aimed at identifying and evaluating causes or risk factors of an outcome event (e.g., fatality or severe crash). "Observation and Experiment" (Rosenbaum, 2017) and "Design of Observational Studies" (Rosenbaum, 2010) are two important textbooks that explain in great detail the characteristics associated with these studies. In an observational study, individual observations in the sample are studied to measure the characteristics of the data population. However, no attempt is made to intervene or influence the variables of interest but the analyst simply observes the data patterns and evaluates the strength of the relationship between the explanatory and dependent variables. By contrast, in an experimental study design, treatment is given to certain individuals or subjects (called as experimental units) and the analyst attempts to isolate the effects of the treatment on the outcome variable. Although the experimental study design identifies the true cause–effect relationship, it is time-consuming and expensive. In addition, in the field of highway safety, it would be considered unethical to treat a site randomly that does not experience any safety issues, which often results in a waste of limited resources and, worse, could lead to additional injuries. Observational studies are widely used in traffic safety analysis, and they primarily include cross-sectional studies, cohort studies, and case–control studies. Of which, the

https://doi.org/10.1016/B978-0-12-816818-9.00002-0

cross-sectional study is the most preferred study approach and is explained in detail in this chapter.

This chapter first describes the different types of data and analysis methods. The discussion includes data and modeling issues and the techniques to overcome them. Then, the chapter presents different functional forms, selection of variables, and the application of the modeling framework. Techniques for determining sample sizes, outlier identification, and transferability of models to other geographical areas are also presented. Lastly, a brief outline of other study designs that are not as commonly used as the cross-sectional study in traffic safety is presented.

6.2 Types of data

Understanding different data types is a critical step before conducting any analysis. This section presents an overview of three types of observational data that are often used in traffic safety analysis.

6.2.1 Time-series data

A series of the same data observations measured and ordered in time is called time-series data. With this type of data, time is usually considered as an independent variable. In traffic safety analysis, time-series data can be used to develop regression models to understand the relationship between traffic crashes occurring over a period and influencing factors that vary over time, such as macroeconomic, socio-demographic, and transport conditions (Quddus, 2018). These models are also used for evaluating the effectiveness of various engineering and nonengineering-based interventions and for forecasting the traffic safety in order to plan and implement appropriate preventive measures (Note that this is different than the methods used in before—after studies described in the next chapter). As the traffic safety outcome data (e.g., traffic fatalities) in the time series analysis are collected at multiple time points at equal intervals, the data are considered to be "interrupted" (Quddus, 2018). When conducting a time-series analysis, structural issues related to internal data, such as autocorrelation, seasonality, and stationarity, need to be addressed or accounted for. They are described as follows:

- *Autocorrelation*: As time-series data are ordered in time, they often display serial dependence. Serial dependence occurs when the occurrence of an event is statistically dependent on the same event that occurred in the past. In such cases, the random errors (i.e., the difference between fitted and observed values) in the model are often positively correlated over time and each random error is more likely to be similar to the previous random error. The time gap between the two correlated observations is called the lag. If the observations are correlated one time interval apart, then it is called lag 1. Similarly, if they are correlated *k* time intervals apart, then it is called lag *k*. Autocorrelation function measures the linear relationship between the lagged values.

- *Seasonality*: If the time series data are affected by seasonal factors such as time of day, day of week, or month of year, then the data series display seasonality. When seasonality exists, the data show short-term movements. There are many graphical techniques to determine the seasonality in the data; among them the autocorrelation plots are typically used. Autocorrelations are calculated and plotted for varying time lags. The seasonality exists if one or more of the autocorrelations are significantly different from zero.
- *Stationarity*: If the statistical properties of time-series data do not change over time, then the data exhibit stationarity. In other words, the data have constant mean and variance, and the covariance is independent of time. Stationarity can be detected by plotting the data or their functions, and determining visually if the property of stationary (or nonstationary) is present in the data.

The autoregressive (AR) model is commonly used for modeling univariate time series data. AR is a linear regression model where the current observation is modeled as a function of one or more lagged values of the series. The moving average (MA) model is another commonly used model where the current observation is modeled as a function of residual errors of one or more lagged values of the series. For each observation, it is assumed that the residual errors are generated from a normal distribution with zero mean and constant variance. The combination of AR and MA models is called the Box—Jenkins ARMA model. The ARMA model assumes that a time series is stationary. The autoregressive integrated moving average (ARIMA) model is a generalization of an ARMA model and is used for nonstationary time series to achieve stationarity by differencing the time series one or more times.

In traffic safety analyses, ARIMA models are commonly used for analyzing time-series safety data and forecasting traffic crashes (Eze and Okonkwo, 2018; Ghédira et al., 2018; Ihueze and Onwurah, 2018). ARIMA models are also used to evaluate the safety effectiveness of various interventions. For instance, these models have been used to quantify the impact of Illinois. Child Passenger Protection Act on fatalities (Rock, 1996) and to examine the improvement in traffic safety after changing of an existing seat-belt law from secondary to primary enforcement (Houston and Richardson Jr, 2002).

For time-series data that exhibit seasonal patterns, a seasonal term is added to the ARIMA model and the model becomes seasonal ARIMA (also called SARIMA). The SARIMA model is also utilized for analyzing the time-series data and predicting traffic safety in the future (Bahadorimonfared et al., 2013; Zhang et al., 2015; Eze and Okonkwo, 2018). Lee et al. (2018) used the SARIMA model to evaluate the effectiveness of marijuana restrictions on the number of marijuana-related crashes after medical legalization.

The limitation with developing the ARIMA or SARIMA model is their requirement of large sample size. The minimum sample size required for developing an accurate SARIMA model is 80 observations (Makridakis et al., 2008). In addition, ARIMA models assume that the data follow a Gaussian distribution. Quddus (2018) pointed out that ARIMA models might be appropriate for highly aggregated time-series data with a large mean but not for highly disaggregated time-series crash count data. Integer-valued autoregressive (INAR) Poisson and the Poisson autoregressive models were introduced to overcome the issues related to time-series crash count data (Brijs et al., 2008; Quddus 2008, 2018; Yannis and Karlaftis, 2010). However, these models have significant difficulties in handling the overdispersion in crash count data (Quddus, 2018). A Negative Binomial Integer-valued Generalized Autoregressive Conditional Heteroscedastic (NBINGARCH) model was proposed to handle the overdispersion in time-series count data (Zhu, 2011) and has been successfully applied in traffic safety to model the relationship between time-series crash counts and influential factors (Ye et al., 2012). Recently, a negative binomial mixture integer-valued GARCH (NBMINGARCH) model has been proposed for modeling time series of counts with the presence of overdispersion because this type of model can handle multimodality and nonstationary components (Diop et al., 2018). Generalized linear autoregressive and moving average (GLARMA) is another model that was applied for modeling time series of crash counts with covariates (Quddus, 2016).

Quddus (2018) developed the recommendations presented in Table 6.1 as an approximate indicator for using an appropriate time-series regression model in analyzing integer-valued crash counts. However, the author

TABLE 6.1 Appropriate regression model for time-series crash count data (Quddus, 2018).

Aggregation level	Sample mean	Recommended model
Highly aggregated	>50	ARIMA
Disaggregated	10–20	Poisson INAR(1), NBINGARCH, or GLARMA
Highly disaggregated	<10	NBINGARCH, or GLARMA

stated that the final selection of the model should be based on appropriate statistical inference, inherent assumptions, and relevant modeling mechanisms, along with the consideration of different model goodness-of-fit (GOF) measures.

6.2.2 Cross-sectional data

Cross-sectional data is a type of observational study data where outcome and exposure are assessed at one point or a short period of time in a sample population. The underlying assumption is that all sites should have similar characteristics (e.g., functional class, traffic control at intersections). Routinely collected data such as crashes, traffic, and geometric data are often used for cross-sectional data analysis, as discussed in Chapter 2—*Fundamentals and Data Collection*. Cross-sectional studies are usually inexpensive and can be conducted relatively faster than time-series studies. Using cross-sectional data, analyses can be conducted on multiple outcomes at the same time. Primarily, there are three advantages of cross-sectional data (Pratt et al., 2018):

- They provide a more robust predictive model than panel data (discussed in Section 6.2.3) when the year-to-year variation in the independent variables is largely random.
- They contain fewer or no observations with missing values, as some operational features may not be collected every year.
- Using cross-sectional data for model calibration minimizes the problems associated with overrepresentation of segments or intersections with zero crashes.

Cross-sectional methods are commonly applied in traffic safety analysis. For instance, crash-frequency models or safety performance functions (SPFs) in the HSM (AASHTO, 2010) are developed using cross-sectional data. However, as mentioned in Mannering (2018), the data in safety analysis are not cross-sectional in the traditional sense because data are not collected at one point of time or space. Instead, crashes are aggregated over a long period (e.g., annually) due to their rarity in occurrence. In addition, no two crashes occur at exactly the same point in space (as discussed in Chapter 3—*Crash-Frequency Modeling*). All crashes that occurred over the length of a highway segment or within an influential area of an intersection are typically considered. Thus, it is important to consider the temporal and spatial components while developing the cross-sectional model. The reader is referred to Chapter 9—*Models for Spatial Data* for methods and techniques for handling spatial data.

In safety modeling, data from a few years are aggregated to develop a cross-sectional model. Study duration (usually in "years") is considered as an offset variable in the regression model (note: the model output is usually in *crashes* per *year*). Bonneson et al. (2012) documented that one of the reasons for preferring cross-sectional data to panel data (discussed in Section 6.2.3) is the accuracy of average annual daily traffic (AADT) in most highway safety databases. After examining states' databases and their documentation, the authors mentioned that the segment AADT volume is frequently extrapolated by the states from partial yearly counts taken at temporary count stations located several miles from the subject segment, which results in less reliable exposure estimates. In addition, when a current count is not available for a segment, states may adjust the AADT volume from last year when it was counted (which could be several years ago) or sometimes, just leave the variable as missing (Bonneson et al., 2012). Consequently, it is common for a segment's AADT volume to be missing for one or more years.

Cross-sectional data do not normally describe which variable is the cause and which one is the effect. This is mainly because data do not include information on confounding factors and other variables that affect the relationship between cause and effect. For the same reason, Hauer (2010) suggested that observational cross-sectional studies cannot be used to draw cause–effect conclusions.

In traffic safety, as discussed earlier, the cross-sectional data include variables collected over some short time period. However, some explanatory variables may change significantly during this time period (e.g., traffic variations in a day, week or month) and are not usually considered due to the lack of detailed information. If these variations are not considered, then potentially important explanatory information may be lost. Due to this unobserved heterogeneity, the model estimates may be biased or unreliable as described in Chapter 3—*Crash-Frequency Modeling* or Chapter 4—*Crash-Severity Modeling*.

6.2.3 Panel data

Panel data, also called longitudinal data, are multidimensional data that include repeated observations of the same variables (e.g., lane width, AADT) over short or long periods of time (i.e., monthly or yearly). Longitudinal study or panel study refers to a study that uses panel data. Time series data that consider one panel member over time and cross-sectional data that consider multiple panel members at one time point can be thought of as special cases of panel data in one dimension only.

Two types of panel datasets exist. The first type is called a balanced panel dataset in which each panel member (e.g., highway segment) is observed every year (as a distinct observation). Consequently, if there are n panel members and t periods, the number of observations in the balanced panel dataset is equal to $n \times t$. The second type is called an unbalanced panel dataset in which at least one panel member (or the characteristics of the panel member) is not observed every period, which is typically the case with traffic safety data (i.e., some variables are not measured in every period). Therefore, if there are n panel members and t periods, then the number of observations in the unbalanced panel dataset is less than $n \times t$.

As stated earlier, data for a longer period (e.g., 3—5 years) are typically considered when developing typical crash-frequency models. Some variables such as retroreflective device performance, friction level, and precipitation rate vary notably between months and years. In such situations, cross-sectional modeling framework may not identify realistic patterns in the variables. Panel data modeling is recommended in such cases when the variables are observed over time. The panel data model allows the safety effects of changing variables to be quantified more precisely when the independent variable value is measured for each site and for each year or month. Panel data modeling has the following advantages (Washington et al., 2011):

- From the statistical perspective, the increase in the number of observations leads to a higher degree of freedom and less collinearity, which in turn improves the parameter estimation accuracy.
- Researchers can test whether or not more simplistic specifications are appropriate. For instance, additional parameters can be introduced into the model to account for cross-sectional heterogeneity. If these parameters are not different from zero, then we can conclude that the cross-sectional homogeneity exists in the data.
- The panel models can be used to analyze some specific questions, such as a change in the variable effect over time, which cannot be answered with cross-sectional modeling.

As discussed by Lord and Persaud (2000), temporal or serial correlation should be considered while analyzing panel data because the same site is repeated multiple times. Random effects models and those estimated using the generalized estimating equations (discussed in detail in Chapter 2—*Fundamentals and Data Collection*) could be used for handling serial correlation in panel data (Lord and Persaud, 2000).

Example 6.1

Using the rural two-lane horizontal curve dataset shown in Pratt et al. (2018), develop cross-sectional data and panel data models and compare the estimated parameters.

This dataset collected in Texas includes 46,753 rural two-lane horizontal curves for a 5-year period. The skid number variable was missing for a few years for some sites and there was an overrepresentation of horizontal curves with zero crash counts. In the cross-sectional data model, the dependent variable used was the sum of crashes over a 5-year period and the independent variables were averaged over the time period. In the panel data modeling, the dependent variable used was the crashes per year. This is an unbalanced panel dataset because some sites were excluded because of missing information for a few variables during some years. For instance, after excluding some observations, the final sample size of the panel dataset is 67,426 observations.

First, determine the functional form:

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

The functional form can be manipulated to get the following after including the significant variables (for more details, the reader is referred to Pratt et al. (2018)):

$$\mu = L \times e^{\beta_0} \times F^{\beta_1} \times \left(1 + \beta_2 (0.147V)^4 \frac{(1.47V)^2}{32.2R^2}\right) \times e^{\beta_3(LW-12)} \times e^{\beta_4(SW-8)}$$
$$\times\, e^{\beta_5(SK-40)} \times e^{\beta_5(AP-30)}$$

In this functional form, L is the length of the horizontal curve, F is the AADT flow, V is the regulatory speed limit, and R is the horizontal curve radius, LW is the average lane width, SW is the average shoulder width, SK is the skid number (a measure of friction levels), and AP is the annual precipitation, respectively.

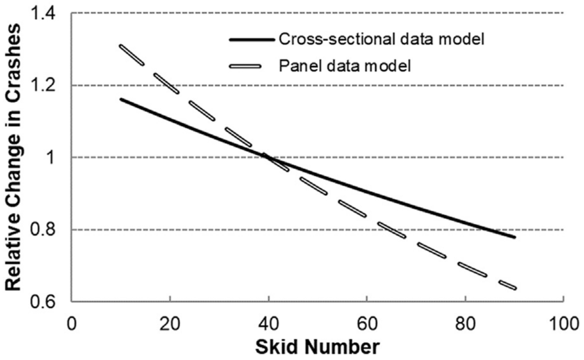
Second, estimate the coefficient of the model using the MLE.

Third, present the results of models.

Example 6.1 (*cont'd*)

Variable	**Cross-sectional data model**			**Panel data model**		
	Estimate	**Std. Error**	***P*-value**	**Estimate**	**Std. Error**	***P*-value**
Intercept (β_0)	−7.733	0.156	<.0001	−7.439	0.246	<.0001
Ln(AADT) (β_1)	0.790	0.019	<.0001	0.760	0.027	<.0001
Curve radius (β_2)	0.461	0.038	<.0001	0.356	0.050	<.0001
Lane width (β_3)	−0.040	0.017	0.0172	−0.064	0.025	0.0094
Shoulder width (β_4)	−0.041	0.006	<.0001	−0.040	0.009	<.0001
Skid number (β_5)	−0.005	0.001	<.0001	−0.009	0.002	<.0001
Annual precipitation (β_6)	0.015	0.002	<.0001	0.014	0.002	<.0001
Dispersion	0.855	0.049	<.0001	0.585	0.073	<.0001
AIC	29,731			16,326		

The following figure shows the comparison of relative change in crashes for the skid number variable from two models. According to the cross-sectional data model, an increase of skid number by 10 units will reduce the crash frequency by 5%. However, as per the panel data model, for the same change in the skid number, the crashes reduce by 9%.



6.3 Data and modeling issues

This section covers important data and modeling issues that should be considered while developing a model. The section includes a description related to over- and underdispersion, low sample mean and small sample size, under reporting, omitted variables bias, endogenous variables, and unobserved heterogeneity.

6.3.1 Overdispersion and underdispersion

As discussed in previous chapters, crash data are usually characterized by overdispersion, which indicates that the sample variance is greater than the sample mean. This is also referred to as unobserved heterogeneity. In such cases, the Poisson model is not an appropriate choice for analyzing crash data. If used, biased and inconsistent parameter estimates may be produced, which in turn could lead to erroneous inferences regarding the factors that determine crash frequencies (Cameron and Trivedi, 2013). In Chapter 3—*Crash-Frequency Modeling*, several models have been proposed, ranging from the most basic but highly popular Poisson-gamma (also called Negative Binomial) to advanced or complex models, such as random-parameters, multivariable (NB-Lindley), and semiparametric models (Dirichlet Process Models). The selection of an appropriate model should be governed by study objectives (e.g., establishing relationships, prediction), the amount dispersion observed in the data, the data generating process (e.g., mixed population, unobserved variables) as well as models that provide a logical choice or goodness-of-logic (Miaou and Lord, 2003). As described in Chapter 3- Crash-Frequency Modeling, selecting a very complex model does not necessarily mean that the model is better, even when the "fit" is superior. The model could be overly complex given the study objectives or the gains they provide compared to traditional models.

On rare occasions, crash data are characterized by underdispersion, where the mean of the crash counts on roadway entities is greater than the variance (also covered in Chapter 3). The observed underdispersion can be attributed to the data generating process or could be the results of the regression analysis where the modeling outcome shows sign of underdispersion (when the outcome is conditional upon the mean). Underdispersion is often observed when the sample mean value is low (which is usually the case with crash data) (Lord and Mannering, 2010). The NB distribution could theoretically handle underdispersion as long as the dispersion parameter, α, is greater than $-1/\mu$ (Saha and Paul, 2005). It should be noted, however, that the inferences cannot be properly estimated when underdispersion occurs, which makes the NB model technically ineffective (see Lord et al. (2008)). To handle under-dispersion, researchers in various fields have proposed alternative distributions or models. Of all the available distributions that have been proposed, three distributions that were used in traffic safety and can handle both over- and underdispersion are the Conway–Maxwell–Poisson (Lord et al. 2008, 2010), the Double-Poisson (Zou et al., 2013), and the Generalized Estimator (Ye et al., 2018) distributions. See Chapter 3—*Crash-Frequency Modeling* for more details.

6.3.2 Low sample mean and small sample size

Due to large costs associated with the data collection process, crash data are often collected at a limited number of sites and this results in a small number of observations in the dataset. In addition, some facilities are unique and rare such as three-legged signalized intersections located in rural areas that result in a small sample size. Along the same line, crash data for some roadway entities may have few observed crashes that result in a preponderance of zero responses. Data characterized by small sample size and low sample-mean can cause estimation problems in traditional count-frequency models. For instance, this kind of data can significantly affect the performance of Poisson-gamma models, particularly the one related to the estimation of the inverse dispersion parameter (Lord, 2006; Lord and Miranda-Moreno, 2008). The resulted goodness-of-fit statistics can also become unreliable when they are estimated using data characterized by small sample size and low sample mean values (Wood, 2002; Park and Lord, 2008). With small sample sizes, the desirable large-sample properties of some parameter-estimation techniques (for e.g., maximum likelihood estimation) are not realized. With low sample means (and a preponderance of zeros), the distribution of crash counts will be skewed excessively toward zero, which can result in incorrectly estimated parameters and erroneous inferences. Lord (2006), Miranda-Moreno et al. (2008) and Ye and Lord (2014) have offered guidelines for determining the minimum sample size to minimize the biases caused by the low sample mean and small sample size (see Section 6.4.6). The data aggregation techniques presented in Section 6.4 can also be used to overcome the small sample size and low sample mean issues when possible.

6.3.3 Underreporting

It has been well-documented that crashes with lower severity levels are less likely to be reported to governmental authorities compared to more severe crashes (Aptel et al., 1999; Elvik and Mysen, 1999). For example, people involved in a reportable property damage only collision (above the minimum reportable threshold) may not be interested in seeing their vehicle insurance premiums go up and would therefore directly pay for the damages themselves or worse, flee from the crash scene (which is more common than we think). Furthermore, there is a possibility of inconsistency in the classification of a crash outcome into no injury or possible injury levels; and/or an arbitrary crash threshold for the vehicle or property damages exceeding a certain amount (Hauer, 2006). These factors make crash data an outcome-based and potentially biased sample. There is a lot of variation in the extent of underreporting, which can depend on the study location and severity levels. For instance, about three decades ago, Hauer and Hakkert (1988) stated that approximately 20% of severe injuries, 50% of minor injuries, and up to 60% of no-injury crashes were not reported. Elvik and Mysen (1999) reported underreporting rates of 30%, 75%, and 90% for serious, slight, and very slight injuries, respectively. According to Blincoe et al. (2002), up to 25% of all minor injuries and almost 50% of no-injury crashes were likely to be nonreported. The underreporting is a more significant issue in low and middle-income countries than in high-income countries. Recent research has shown that crash-frequency and crash-severity models are likely to produce biased estimates when underreporting is not considered in the model-estimation process. Some studies have proposed methods to minimize this bias even if the underreporting rate is unknown (see Kumara and Chin (2005); Yamamoto et al. (2008); Ma (2009) Ye and Lord (2011); Patil et al. (2012)).

6.3.4 Omitted variables bias

Omitted variable bias occurs when important explanatory variables that are correlated with the dependent variable are not included in the model. The amount of bias is a function of the correlation strength. Leaving out important explanatory variables results in biased parameter estimates that can produce erroneous inferences and crash-frequency forecasts (see Washington et al. (2020)). Explanatory variables that are omitted from the model cause confounding effect and are thus known as confounding variables or confounders. The results of omitted variables lead to the over- or underestimation of the strength of an effect, change the sign of an effect, and may mask the actual effect on the dependent variable (Wu et al., 2015). If the omitted variable is positively (or negatively) correlated to the included explanatory variable and the dependent variable, then there is a positive bias and the parameter is overestimated.

Alternatively, if the omitted variable is positively (or negatively) correlated to the included explanatory variable and negatively (or positively) correlated to the dependent variable, then there is a negative bias and the parameter is underestimated. Therefore, it is important to run correlation statistics before the start of model development. Chapter 5—*Exploratory Analyses of Safety Data* presents the details on how to run the correlation statistics. In situations where we suspect that the outcome variable depends on unobservable explanatory variables correlated with the observed explanatory variables, panel data are most useful. Panel data estimators consistently estimate the effect of the observed explanatory variables if omitted variables are constant over time.

6.3.5 Endogenous variables

An endogenous variable is an explanatory variable whose value is determined or influenced by one or more variables in the model. In contrast, exogenous variable values are not determined or affected by changes in the other variables of the model. Carson and Mannering (2001) studied the endogeneity problem by exploring the effectiveness of ice-warning signs in reducing the frequency of ice-related crashes. An indicator variable for the presence of an ice warning sign is typically used when developing a crash-frequency model. As ice-warning signs are more likely to be placed at locations with high numbers of ice-related crashes, this indicator variable may be endogenous (the explanatory variable will change as the dependent variable changes). If this endogeneity is ignored, the parameter estimates will be biased. In the case of the ice-warning sign indicator, ignoring the endogeneity may lead to the erroneous conclusion that ice-warning signs actually increase the frequency of ice-related crashes because the signs are going to be associated with locations of high ice-crash frequencies. Kim and Washington (2006) studied the effectiveness of left-turn lanes at intersections on the angle crashes. Left-turn lane is considered an endogenous variable because it is more likely to be placed at intersections with a high number of left-turn related crashes. To address the endogeneity problem, Kim and Washington (2006) used a limited-information maximum likelihood estimation approach. First, they developed a logistic regression model with a left-turn lane presence indicator as a dependent variable and angle crash frequency as one of the explanatory variables. The estimation results for the left-turn lane indicator were then used as an explanatory variable in the crash count model. Their study results showed that left-turn lanes increase crashes when not accounting for endogeneity; however, this variable showed a negative effect on angle crash frequencies when endogeneity is considered in the model.

6.3.6 Unobserved heterogeneity

Traditionally, fixed effects models have been used to evaluate the effect of the explanatory variables on the frequency of crashes. As discussed in Chapter 3—*Crash-Frequency Modeling*, it is assumed that there is only one source of random variation and this variation is from the random sampling process. However, it is often found that there are unobserved variations from one roadway segment to the next (unobserved heterogeneity), especially when panel data are considered. In such cases, random effects models are recommended. In the random effects model, a term is added to the model that captures the variation from one site to another; however, the effect of explanatory variables is assumed to be the same among sites. The site-specific effect is a random variable that is uncorrelated with the explanatory variables and assumed to have a constant (nonzero) variance. The Hausman test (Hausman, 1978) can be used to determine whether a random effects or fixed effects model is appropriate. Hausman et al. (1984) first examined the random effects negative binomial model in the context of panel count data to analyze the relationship between patents and R&D expenditures.

In addition, traditional statistical modeling does not allow parameter estimates to vary across observations. This implies that the effect of the explanatory variable on the frequency of crashes is constrained to be the same for all observations (for e.g., the effect of traffic control devices is the same for all horizontal curves in the database). However, because of unobserved variations from one roadway segment to the next (unobserved heterogeneity), one might expect the estimated parameters of some explanatory variables to differ across roadway segments. If some parameters do vary across observations, then it is called the random parameters model (RP). The random effects model is a special case of the random parameters model in which only the intercept variable is considered random. The estimated parameters are assumed to be normally distributed with mean zero and a specified variance. Estimation techniques exist for allowing parameters to vary across observations, but the model estimation process becomes considerably complex. These techniques are described in Chapter 3—*Crash-Frequency Modeling*.

Although the random parameters models capture the unobserved heterogeneity via the parameters, multiparameter models, such as the NB-Lindley (NB-L) and NB-Generalized Estimate (NB-GE), reduce the unobserved heterogeneity by minimizing model errors. The errors are reduced as these models have two or more "shape" parameters which offer more flexibility than the traditional the NB model to "fit" the data. The process for selecting the models between the RP models and multiparameter models should be based on the study objectives, the characteristics of the data and the method selected for estimating the parameter (maximum likelihood estimate vs. Bayesian methods). Additional information on this topic can also be found in Chapter 3—*Crash-Frequency Modeling*.

6.4 Data aggregation

Data used for safety analyses have unique characteristics that are not typically found in other disciplines. The important characteristic is related to datasets that include a large amount of zero responses. As documented in Lord and Geedipally (2018), excess zero observations are often attributed to how data are assembled or formatted on spatial or temporal scales. For example, it is expected to see more zero observations in data that are aggregated weekly than monthly or yearly. Crash data at a site are usually defined as a count number over the space and time. Therefore, the number of zero observations in the compiled dataset is directly correlated with the selected spatial and/or temporal scales. By adjusting the time and spatial scales, the number of zero responses observed in the dataset can increase or decrease. For example, by changing the segment length of a site from 0.1 mile to 1 mile, the number of zero observations in the complied dataset will be reduced as the new segment will include all of the crashes on the segments now aggregated. Similarly, changing the time scale from monthly durations to yearly periods will result in a reduction of the number of zero responses in the dataset. This is depicted in Fig. 6.1.

The root of having excess zero observations in data can be described by four major factors (Lord et al., 2005):

- using spatial or time scales that are too small,
- under or miss-reporting of the number or severity of crashes,
- sites characterized by low exposure and high risk; and,
- bias caused by omitting of important variables in the crash data process (i.e., lowers the long-term mean).

The first factor can potentially be overcome by adjusting the time and scale while compiling the datasets (Lord and Geedipally, 2018). Finding a balance in aggregation is a critical task in data preparation. On the one hand, using the disaggregated data may result in having excessive zero

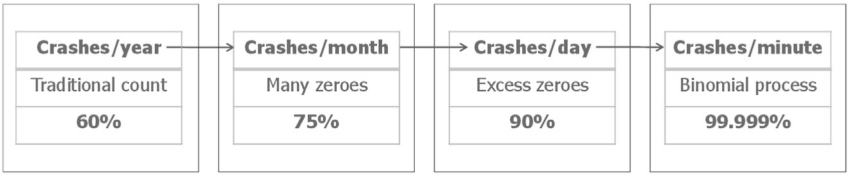


FIGURE 6.1 Percentage of zero responses when changing the time scale (Lord and Geedipally, 2018).

observations, in which the traditional negative binomial (NB) model may not be appropriate for the safety analysis (Lord and Geedipally, 2018). On the other hand, too much aggregation may result in loss of information (Usman et al., 2011), although it may make the NB model a better alternative. The latter characteristic is also called "aggregation bias" or "ecological fallacy" (see, Davis (2004)). Recently, Shirazi et al. (2020) addressed the issue related to aggregated and disaggregated data by conducting a simulation study and measuring the information loss as a function of the precision or accuracy in the coefficient estimation. The authors recommended the following conservative criteria:

- When the percentage of zeros is higher than 70%, aggregate the data only if the change in Coefficient of Variation (CV) of all exploratory variables is less than 10% between aggregated and disaggregated datasets.
- When the percentage of zeros is less than 70%, aggregate the data only if the change in CV of all exploratory variables is less than 5% between aggregated and disaggregated datasets.

When the adjacent segments are aggregated, homogeneity is lost. In such cases, a variable that shows the proportion of presence of a particular roadway characteristic can be adopted (e.g., instead of using a variable that shows whether the horizontal curve is present or not, we can use the proportion of segment length with the presence of horizontal curves).

6.5 Application of crash-frequency and crash-severity models

This section discusses the application of crash-frequency and crash-severity models in highway safety. It first covers different functional forms when entering a variable into the model, and modeling framework. Then, the discussion includes variable selection, crash variance and confidence interval estimation, sample size determination, outlier analysis, and model transferability.

6.5.1 Functional form

Most crash-frequency models assume that explanatory or independent variables influence the dependent variables in some linear manner (more specifically, log-linear relationship is often adopted); however, there is no logical reason for this assumption, except for simplicity. It is in fact argued that the simple log-linear structure may not truly represent the complexity of the process by which variables combine to cause crashes (Hauer, 2015; Wu and Lord, 2017). There is a body of work that suggests that nonlinear functions better characterize the relationships between crash frequencies and explanatory variables. These nonlinear functions

TABLE 6.2 Functional form for different variables (Hauer, 2015).

Exposure variables	Influential variables
1 Power: X^{β_1}	**5** Exponential: $e^{\beta_1 X}$
2 Polynomial: $\beta_1 X + \beta_2 X^2 + \beta_3 X^3 \ldots$	**6** Linear: $1 + \beta_1 X$
3 Logistic: $1/(1 + \beta_1 e^{\beta_2 X}) - 1/(1 + \beta_1)$	**7** Quadratic: $1 + \beta_1 X + \beta_2 X^2$
4 Weibull: $1 - e^{-(X/\beta_1)^{\beta_2}}$	

can often be quite complex and may require estimation procedures (i.e., Bayesian methods) (Miaou and Lord, 2003). Hauer (2015) presented a sample of four different types of functions for exposure variables (such as segment length and AADT) and these functions start at the origin and take positive values only. He also presented three types of functions for other explanatory variables (such as lane width or shoulder width) and these functions start with the ordinate of 1. The functions are shown in Table 6.2.

A variable can be represented in multiple ways in the same model. If it is a continuous variable, then it can be entered into the model either as a continuous or a categorical variable after dividing the data into categories. Another way is to fit separate functions for different categories. For example, Bonneson et al. (2012) used the following functional form to represent the effect of lane width on freeway crashes.

$$CMF_{lw} = \left[\begin{array}{ll} e^{b_w(W_l - 12)} & \text{If } W_l < 13\,ft \\ e^{b_w(1.0)} & \text{If } W_l \geq 13\,ft \end{array}\right. \tag{6.1}$$

where CMF_{lw} is the lane width crash modification factor, W_l is the lane width, and b_w is the parameter to be estimated.

Crash modification factors (CMFs) are factors that are associated with the predictive methodology documented in the Highway Safety Manual (HSM) (AASHTO, 2010). These multiplicative factors adjust the estimated values produced from baseline crash-frequency models (called as safety performance function or SPFs in the HSM) to describe changes in operational and design characteristics (such as the shoulder width changing from 2 to 8 ft). A value above 1 indicates an increase in crashes and a value below 1 indicates a reduction in crashes. CMFs are commonly developed from before-after studies (discussed in detail in Chapter 7—*Before-After Studies in Safety*) because they are considered to be more reliable. However, there are many instances where the CMFs are developed from cross-sectional studies.

In a regression model, the independent variables might interact with each other. In such cases, the relationship between the dependent and an independent variable is influenced by a third variable. Jaccard and Turrisi (2003) called the third variable a "moderator" variable and the interaction is referred to as a "moderated" causal relationship. Although the interaction effect makes the model more complex, it is important to incorporate this effect in the model in order to better understand the underlying mechanism. As pointed out by Hauer (2015, page 188), "a model equation that is a product of single variable factors cannot accommodate interaction and cannot be a source of CMFs that are a function of other variables." If interactions are present, Hauer (2015) recommended that the variable be a function of the relevant moderator variables to draw meaningful conclusions. In the safety literature, it is not common to find the interaction effects in the SPFs, mainly because of the difficulty to understand which interactions are important, and how they should be included in a model, unless there is a theoretical reason for including them (Srinivasan and Bauer, 2013). However, there are many instances where an interaction is included in the SPFs. The most common interaction variables used in the SPFs are curve radius and length of the curve (e.g., see Bauer and Harwood (2013); Bonneson et al. (2012)).

The ideal way to decide which functional form to use for a variable is to develop a scatterplot (these plots are explained in detail in Chapter 5—*Exploratory Analyses of Safety Data*) with the variable of interest and the crashes (after accounting for exposure). Fig. 6.2 shows the scatterplot of intersection crashes on vertical axis and cross street entering volumes on the horizontal axis. As traffic volume is an exposure variable, functions in the first row of Table 6.2 can be used. The solid line in Fig. 6.2 shows a fit using the power function.

Cumulative residual (CURE) plots, described in detail in Chapter 2—*Fundamentals and Data Collection*, are used to obtain further insight into whether the selected appropriate functional form was reasonable

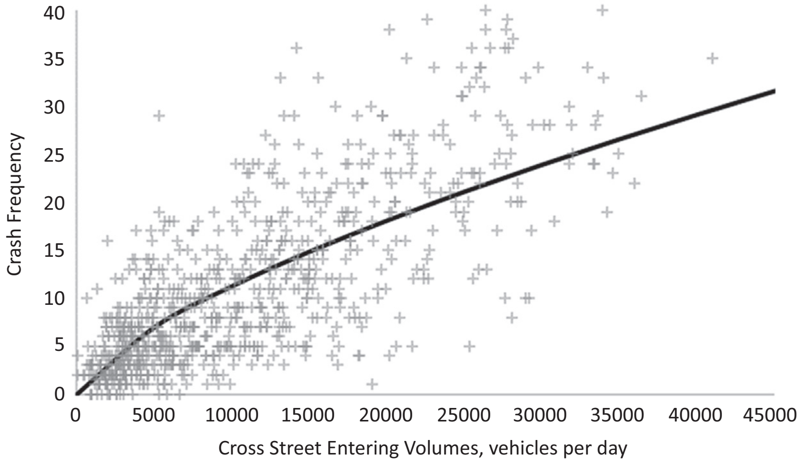


FIGURE 6.2 Relationship between cross street entering volumes and crash frequency.

(Hauer, 2004). This plot contains the variable values on the horizontal axis and cumulative residuals with their confidence intervals on the vertical axis. To create these plots, the residuals are first calculated by considering the difference between the observed number of crashes and the predicted number of crashes from the crash-frequency model. Then, the observations are sorted in increasing order of the variable and the plot is generated. If the residuals are within the interval, then it can be concluded that the functional form for the variable is appropriate.

The inclusion criteria of variables into the model equation should not only base on the GOF measures but also on the logic (e.g., reason, consistency, and coherency), flexibility, extensibility, and interpretability of the functional form. Miaou and Lord (2003) called this concept as "goodness-of-logic" (GOL). The authors mentioned that we have plenty of GOF measures from statistics, but the GOL measures from engineering and other perspectives are almost nonexistent. In general, imposing logical constraints on the functional form reduces the "solution space" and thus, decreases the achievable GOF of the resulting model. It can, however, enrich the logical interpretation of the functional form, complement the limitation of data in size and coverage, and potentially allow the estimated response surface to be more extensible beyond the data range (Miaou and Lord, 2003). Using the system engineering language, logical constraints increase "observability" and enable us to better estimate the true state of the system under study.

Crash-frequency models relate the site crash frequency to their traffic, geometric, and environmental characteristics. There are at least three different ways in which crash-frequency models can be used by agencies to make informed safety decisions (AASHTO, 2010). One application is to use crash-frequency models (although the HSM uses the terminology SPFs, we continue using the term crash-frequency model or CFM for the rest of this section) as part of network screening to identify sections that may have the best potential for improvements (see Chapter 8—*Identification of Hazardous Sites* about how CFMs are used in network screening). The second application is to use CFMs to determine the safety impacts of design changes at the project level. The third application is to use as part of a before-after study to evaluate the safety effects of engineering treatments (see Chapter 7—*Before-After Studies in Safety* about how CFMs are used in before-after evaluation). However, it is important to remember that the regression models can be accurately used for predicting the expected total crash experience for a location, but they may not provide satisfactory results in isolating the effects of individual geometric or traffic control features (Harwood et al., 2000). Three kinds of CFMs have been described in highway safety literature.

6.5.1.1 Flow-only models

The first method consists of developing a model with flow-only variables for average conditions found in the data for each transportation element. These models, sometimes called *general flow-only* models, can be used for cases when limited information about the geometric design features is available. They can still be useful and provide an average value for the safety performance of highway segments or intersections. Although these models suffer from omitted-variable bias, they are typically used in the network screening process to identify hazardous sites (see Chapter 8—*Identification of Hazardous Sites* for more details). The form of the model for roadway segments is as follows:

$$\mu_{rs} = \beta_0 \times L \times AADT^{\beta_1} \tag{6.2}$$

where μ_{rs} is the predicted crashes on the roadway segment; $AADT$ is the annual average daily traffic volume, L is the segment length and β_i are the regression coefficients ($i = 0, 1 \ldots$). This form assumes that the segment length has a linear effect on crashes (i.e., segment length is considered as an offset instead of a covariate). This means that, for example, a 2-mile road section experiences exactly twice the number of crashes as that of a 1-mile section, if everything else remains constant. However, some of the previous studies considered segment length as a covariate as well. When the segment length is a utilized as a covariate, it is implicitly used to capture the factors not included in the model. In such situations, the form of the model is as follows:

$$\mu_{rs} = \beta_0 \times L^{\beta_1} \times AADT^{\beta_2} \tag{6.3}$$

The most common form of the model used for intersections is as follows:

$$\mu_{int} = \beta_0 \times AADT_{maj}^{\beta_1} \times AADT_{min}^{\beta_2} \tag{6.4}$$

where μ_{int} is the predicted crashes on roadway segment; $AADT_{maj}$ = entering annual average daily traffic volume on the major approach of the intersection; $AADT_{min}$ is the entering annual average daily traffic volume on the minor approach of the intersection; and, β_i are the regression coefficients ($i = 0, 1, \ldots$). (See Fig. 2.2 in Chapter 2—*Fundamentals and Data Collection* for an explanation about how the entering flows are calculated.). Miaou and Lord (2003) proposed alternative functional forms that could be used when modeling intersection crashes. These functional forms are as follows.

$$\mu_{int} = \beta_0 \times (AADT_{maj} + AADT_{min})^{\beta_1} \tag{6.5}$$

$$\mu_{int} = \beta_0 \times (AADT_{maj} \times AADT_{min})^{\beta_1} \tag{6.6}$$

$$\mu_{int} = \beta_0 \times (AADT_{maj} + AADT_{min})^{\beta_1} \times \left(\frac{AADT_{min}}{AADT_{maj}}\right)^{\beta_2} \tag{6.7}$$

$$\mu_{int} = \beta_0 \times AADT_{maj}^{\beta_1} \times AADT_{min}^{\beta_2} \times e^{\beta_3 \times AADT_{min}} \tag{6.8}$$

$$\mu_{int} = \left(AADT_{maj} \times e^{\beta_0 + \beta_1 \times AADT_{min}}\right) + \left(AADT_{min} \times e^{\beta_2 + \beta_3 \times AADT_{maj}}\right) \tag{6.9}$$

It should be pointed out that Eq. (6.9) can only be estimated using the Bayesian method, as it contains two mean values. This functional form should be used if the boundary conditions are critical in the analysis of the safety performance of intersections. Normally, all general flow-only models documented in safety literature are estimated using the Poisson-gamma model.

6.5.1.2 Flow-only models with CMFs

With this method, the models are developed using only data that represent a given set of baseline conditions, as opposed to the general flow-only models described in the previous section. The baseline conditions usually reflect the nominal conditions agencies most often used for designing segments and intersections (e.g., 12-ft lanes and 8-ft shoulders). The base condition model is calibrated using a database that is assembled to include only segments or intersections that have characteristics equal to base conditions, and it accounts for exposure to traffic flow as the only independent variable (similar to the functional form found in Eqs. 6.3–6.9). With these types of models, changes in geometric design characteristics are estimated using CMFs. For this modeling structure, the base models and CMFs are estimated independently. The structure of the crash prediction algorithm is as follows:

$$\mu = \mu_b \times (CMF_1 \times CMF_2 \times \ldots .CMF_n) \tag{6.10}$$

where μ is predicted crashes of an entity; μ_b is the predicted crashes for base conditions (note that the functional forms presented in Eqs. (6.3–6.9) are used to develop base models), and CMF_1 CMF_2, and CMF_n are crash modification factors for various features (1, 2, …, n). It should be pointed out that the uncertainty associated with the estimated or predicted value increases significantly as the number of CMFs is used to adjust the predicted value (Lord, 2008). These models are normally estimated using the Poisson-gamma model.

Exercise 6.1

Using the Texas rural multilane undivided highway dataset, develop flow-only models for average and baseline conditions.

continued

Exercise 6.1 (*cont'd*)

First, determine the baseline conditions. For simplicity, let us just consider lane width, shoulder width, and number of horizontal curves variables in the baseline conditions. The most common values in this dataset for these three variables are 12-ft lanes, 8-ft shoulders, and no horizontal curves. After excluding the observations that do not satisfy these conditions, the number of observations in the dataset reduced from 1164 to 138.

Second, estimate the coefficient of the model using the Poisson-gamma modeling structure.

Third, present the results of models.

	Flow-only Model for Average Conditions			**Flow-only Model for Baseline Conditions**		
Variable	**Estimate**	**Std. Error**	***P*-value**	**Estimate**	**Std. Error**	***P*-value**
Intercept (β_0)	−8.605	0.433	<.0001	−10.635	1.332	<.0001
Ln(AADT) (β_1)	1.112	0.050	<.0001	1.358	0.157	<.0001
Dispersion	0.768	0.042	<.0001	0.836	0.150	<.0001
AIC	6362			599		

The result shows that there is a considerable difference between the estimated parameters. The standard errors increased significantly for the baseline model, mainly due to the small sample size. It should be noted that the Akaike Information Criterion (AIC) values cannot be compared between the models because the sample size is different.

6.5.1.3 Model with covariates

With this method, the crash-frequency model is estimated using a database within which each safety-related variable (e.g., lane width, median width) has a representative range of values. Each variable is included in the model and their coefficients are calibrated using regression analysis. These models can be called as a multiple regression model.

All the models described in Chapter 3—*Crash-Frequency Modeling* and Chapter 4—*Crash-Severity Modeling* would be applicable here. The form of crash-frequency prediction algorithm for roadway segments, for example, can be defined as follows:

$$\mu = \beta_0 \times L \times AADT^{\beta_1} \times \exp(\beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n) \tag{6.11}$$

where x_k is the safety-related variable (e.g., lane width, median width, turning lane) ($k = 0, 1, 2 \ldots.$). β_j are the regression coefficients ($j = 0, 1, 2 \ldots.$). Note that this functional form shows a simple log-linear relationship. However, x_i can also represent nonlinear relationships and interactions (see Section 6.5.1 for more details). The full models can be used for estimating base models (by replacing the values as per base conditions) and CMFs simultaneously. This modeling procedure has the advantage of overcoming the regression-to-the-mean (RTM) bias, as the models and CMFs are estimated using cross-sectional data and RTM only affects before–after studies. This is even more important given recent research that showed that the empirical Bayes (EB) method can still provide a biased estimate when it is used for developing CMFs using before–after data (Lord and Kuo, 2012). More discussion of this issue is provided in Chapter 7—*Before–After Studies in Safety.*

6.5.2 Variable selection

This section only applies to models with covariates. Stepwise regression is a data mining tool that is used to build a regression model by selecting the explanatory variables based on their statistical significance. From a group of candidate explanatory variables, the variables are added or removed, one by one, for a multiple regression model through the iterative process, typically using the *P*-values (note: $\mathrm{P} - \mathrm{value} = P(Z \leq z)$ where $\mathrm{z} - \mathrm{statistics} = \mathrm{Coefficient\ Estimates/Standard\ Error}$). The variables are removed if the *P*-values are greater than a prespecified critical value. The most common cut-off critical value considered in the literature is 0.05 but higher values can also be used. It should be noted that the *P*-values used should not be treated too literally. The judgment for including or excluding of any variable should not be completely based on the *P*-values. As such, the GOL concept discussed in the previous section should be implemented always. Three types of selection rules are discussed next.

Forward selection rule starts with no explanatory variables in the model. Variables are added, one by one, based on which variable has the lowest *P*-value. The process is repeated until there are no remaining statistically significant variables and no improvement in the goodness-of fit (GOF) measures such as log-likelihood ratio (LR), the AIC and the

Bayes Information Criterion (BIC) can be observed (see Chapter 2—*Fundamentals and Data Collection for details about the GOF measures*).

Backward elimination rule is the simplest of all variable selection procedures and starts with all possible explanatory variables into the model. The variable with the highest P-value is discarded and the model is refitted. The discarding continues until each variable remaining in the model is statistically significant. Backward elimination is challenging if there is a large number of explanatory variables and it cannot be used if the number of explanatory variables is larger than the number of observations.

Bidirectional elimination is the combination of forward and backward elimination procedures. The procedure starts with no explanatory variables in the model and adds variables using the lowest P-values. In each iteration, the procedure includes a new significant variable in the model but also considers the statistical consequences of dropping variables that are previously included. The process is repeated until there are no remaining statistically significant variables and there is no improvement in the goodness-of fit measures such as the LR and AIC. All the commercial statistical packages have built-in functions that run these different variable selection procedures automatically.

Exercise 6.2

Using the dataset shown in Exercise 6.1, select the variables using the backward elimination rule and 5% significance level.

The dataset includes seven independent variables. Before developing the model, it is important to decide how to include a variable that can provide a logical interpretation. For example, the number of horizontal curves or minor intersections on a segment may not provide a meaningful explanation because segments are of varying length. Instead, we can consider the density (e.g., minor intersections per mile) by accounting for the segment length.

In the first iteration, we will include all seven variables. Two variables have a P-value greater than 0.05 (i.e., above 5% significance level). The P-value for the railroad crossing presence is the highest so we will remove it. In the second iteration, we will run the model without the railroad crossing presence variable. Now, only one variable has a P-value greater than 0.05. In the last iteration, we will exclude the lane width variable and rerun the model.

Exercise 6.2 (*cont'd*)

	First Iteration			Second Iteration		
Variable	**Estimate**	**Std. Error**	***P*-value**	**Estimate**	**Std. Error**	***P*-value**
Intercept (β_0)	−6.029	0.723	<.0001	−6.036	0.723	<.0001
Ln(AADT) (β_1)	1.047	0.046	<.0001	1.047	0.046	<.0001
Lane width (β_2)	−0.074	0.044	0.0922	−0.073	0.044	0.0949
Shoulder width (β_3)	−0.029	0.007	<.0001	−0.029	0.007	<.0001
Railroad crossing presence (β_4)	−0.045	0.158	0.7763			
Posted speed limit (β_5)	−0.024	0.003	<.0001	−0.024	0.003	<.0001
Minor intersection density (β_6)	0.037	0.005	<.0001	0.037	0.005	<.0001
Horizontal curve density (β_7)	0.026	0.012	0.026	0.026	0.012	0.0262
Dispersion	0.600	0.036	<.0001	0.600	0.036	<.0001
AIC	6187			6185		

	Third Iteration		
Variable	**Estimate**	**Std. Error**	***P*-value**
Intercept (β_0)	−6.964	0.464	<.0001
Ln(AADT) (β_1)	1.054	0.046	<.0001
Lane width (β_2)			

Continued

continued

Exercise 6.2 (*cont'd*)

—cont'd

	Third Iteration		
Variable	**Estimate**	**Std. Error**	***P*-value**
Shoulder width (β_3)	−0.031	0.007	<.0001
Railroad crossing presence (β_4)			
Posted speed limit (β_5)	−0.024	0.003	<.0001
Minor intersection density (β_6)	0.036	0.005	<.0001
Horizontal curve density (β_7)	0.027	0.012	0.0247
Dispersion	0.603	0.036	
AIC	6186		

Note that there is no improvement after removing the lane width variable. Although not significant at the 5% significance level, the lane width variable is intuitive and within the logical boundaries and deserves to be included in the model. For this reason, it is important that the *P*-value should not be considered literally. As such the GOL phenomenon discussed earlier should be adopted.

Stepwise regression has certain drawbacks. The main limitation is that some real explanatory variables that could have a strong association with the dependent variable may not be statistically significant, while some irrelevant variables may be coincidentally significant. As a result, the "optimal" model may not be selected and as such the model may fit the data well, but predict the results using a new dataset poorly. In addition, the statistical significance is correlated with the data variability. Even though the variable may influence the dependent variable but *P*-value could be much higher if there is low variation in the data. Variables that are dropped from the model can still be correlated with the dependent variable. Thus, it does not necessarily mean that the variables are unrelated to the response, but they just provide no additional explanatory effect beyond those variables already included in the model.

6.5.3 Crash variance and confidence intervals

As discussed in the previous chapters, the Poisson-gamma (or NB) is a widely used framework in modeling traffic crashes. The crash variance is assumed to be a simple function of crash mean and is estimated as follows.

$$Var(y) = \mu + \alpha\mu^2 \tag{6.12}$$

where μ is the mean of the observation; $Var(y)$ is the variance of the observed crashes, and α is defined as the dispersion parameter of the Poisson-gamma model. Recall that if $\alpha \rightarrow 0$, then the crash variance equals the crash mean and this model reduces to the standard Poisson regression model. Other variance functions also exist for the Poisson-gamma model, but they are seldom used in highway safety studies. The reader is referred to Cameron and Trivedi (2013) for a description of alternative variance functions.

It is argued that the simple variance function cannot explicitly explain the heterogeneity in the sample data, especially when important variables are not observed, cannot be observed, or omitted from the model. The dispersion parameter plays a critical role in estimating the crash variance. Note that the dispersion parameter or its inverse plays an important role in safety analyses, including the computation of the weight factor for the EB method, the estimation of confidence intervals around the gamma mean, and the predicted values of models.

In many earlier studies, the dispersion parameter of Poisson-gamma models was assumed to be invariant of the characteristics of the observations under study. Hauer (2001) first pointed out that the dispersion parameter of Poisson-gamma models should not be fixed, and should be dependent upon the length of the highway segment. He proposed two parametrizations as shown in the following, but noted that other parametrizations may be possible.

$$\alpha_i = e^{\gamma_0} L_i^{\gamma_1} \tag{6.13}$$

$$\alpha_i = \frac{1}{e^{\gamma_0} L_i} \tag{6.14}$$

where α_i is the dispersion parameter for site i, L_i is the segment length, and γ_j are the regression coefficients ($j = 0, 1, 2 \ldots$). Eq. (6.14) is used in the predictive methodology of the HSM.

Miaou and Lord (2003) also noted that the dispersion parameter can be dependent upon the intersection entering flows of crash-flow predictive models, suggesting that the variance function has an unobserved structure. They hypothesized that, due to complexity and interaction of traffic flows near intersections, the unmodeled heterogeneity of the mean of crash counts could be spatially structured. The authors used the following functional form:

$$\alpha_i = e^{\gamma_0 + \gamma_1 \times AADT_{maj,i} + \gamma_2 \times AADT_{min,i} + \gamma_3 \times AADT_{min},i / AADT_{maj,i}} \tag{6.15}$$

Mitra and Washington (2007) confirmed the results of Miaou and Lord (2003) and indicated that a model with a mis-specified mean function will make the variance function dependent upon the covariates of the models. They investigated four different functional forms and concluded that the varying dispersion parameter may not be needed when the functional form describing the mean function is fully specified. Geedipally et al. (2009) examined empirically whether the dispersion parameter should solely be characterized using the length of the segment, as proposed by Hauer (2001). The authors investigated alternative parametrizations, including the ones proposed by Hauer (2001), that would offer a better approach to characterizing the variance function of the Poisson-gamma model as a function of length and/or traffic flow. The results of their study showed that there is no single functional form or parametrization that is suitable for all datasets. Traffic flow was more significantly associated with the variation in the data than segment length. It is therefore recommended that transportation safety analysts evaluate different parametrizations and select the most appropriate one using a combination of GOF criteria, including the significance of the model's coefficients (Geedipally et al., 2009).

Confidence intervals (CIs) and prediction intervals (PIs) can be used for selecting highway design alternatives where the safety performance is used as a screening criterion and for identifying hazardous sites. Wood (2005) proposed a method for estimating the PIs for the predicted response (i.e., crash frequency at a new site, y_i that has similar characteristics as the sites used in the original dataset from which the model was developed) and the gamma mean (m_i), as well as CIs for the true mean crash frequency (alternately called as the mean response or Poisson mean, μ_i), from the NB (Poisson-gamma) regression model. It is important to note the distinction between the Poisson parameter and the Poisson mean. In the case of standard Poisson regression, the two values are in fact equal. However, in a mixed-Poisson model, introduction of an error term into the Poisson parameter makes it such that the two terms are no longer equal. Table 6.3 gives the equations for calculating the confidence intervals. In this table, η is the logarithm of the estimated mean response μ,

TABLE 6.3 Confidence intervals for mean response, gamma mean, and predicted response (Wood, 2005).

Parameter	Intervals
μ	$\left[\dfrac{\overset{\text{I}}{\mu}}{e^{1.96\sqrt{Var\left(\overset{\text{I}}{\eta}\right)}}}, \widehat{\mu}e^{1.96\sqrt{Var\left(\overset{\text{I}}{\eta}\right)}}\right]$
m	$\left[\max\left\{0, \widehat{\mu} - 1.96\sqrt{\widehat{\mu}^2\text{var}\left(\widehat{\eta}\right) + \dfrac{\overset{\text{I}}{\mu}^2\text{var}\left(\overset{\text{I}}{\eta}\right) + \overset{\text{I}}{\mu}^2}{\varphi}}\right\}\right.,$ $\left.\widehat{\mu} + 1.96\sqrt{\widehat{\mu}^2\text{var}\left(\widehat{\eta}\right) + \dfrac{\overset{\text{I}}{\mu}^2\text{var}\left(\overset{\text{I}}{\eta}\right) + \overset{\text{I}}{\mu}^2}{\varphi}}\right]$
Y	$\left[0, \left\lfloor \widehat{\mu} + \sqrt{19}\sqrt{\widehat{\mu}^2 Var\left(\widehat{\eta}\right) + \dfrac{\overset{\text{I}}{\mu}^2 Var\left(\overset{\text{I}}{\eta}\right) + \overset{\text{I}}{\mu}^2}{\phi}} + \widehat{\mu}\right\rfloor\right]$

Note: $Var\left(\widehat{\eta}\right) = XI^{-1}X^T$ where I^{-1} is the variance-covariance matrix and X is a matrix containing observed values in logarithmic form. $\lfloor x \rfloor$ denotes the largest integer less or equal than x.

while φ is the inverse dispersion parameter. Ash et al. (2020) recently expanded on the work by Wood (2005) to include CIs and PIs for several other models, such as the Poisson-lognormal, Poisson-Inverse Gaussian, Poisson—Weibull and Sichel (SI).

6.5.4 Sample size determination

Enough data need to be collected for developing reliable NB regression models. As discussed in Lord (2006), the sample size is governed by the characteristics of the sample mean of the data. The recommended sample size (number of sites) is summarized in Table 6.4.

For the Full Bayes method, the minimum sample size is described in Table 6.5. The minimum sample size was initially developed for Poisson-Lognormal models, but is also applicabe for other mixed-Poisson models.

Ye and Lord (2014) examined and quantified the effects of different sample sizes on the performance of the three most commonly used crash severity models: the multinomial logit, ordered probit, and mixed logit models. The authors recommended the absolute minimum numbers of observations for the ordered probit, multinomial logit, and mixed logit models to be 1000, 2000, and 5000, respectively.

TABLE 6.4 Recommended sample size (Lord, 2006).

Population sample mean	Minimum sample size
5.00	200
4.00	250
3.00	335
2.00	500
1.00	1000
0.75	1335
0.50	2000
0.25	4000

TABLE 6.5 Recommended minimum sample size for Bayesian Poisson-lognormal models (Miranda-Moreno et al., 2008).

Population sample mean	Minimum sample size
≥2.00	20
1.00	100
0.75	500
0.50	1000
0.25	3000

6.5.5 Outlier analysis

An outlier in the data can significantly influence the parameter estimates in the crash-frequency model and consequently predict incorrect crash counts. The outlier techniques such as box-whisker plots and scatterplots presented in Chapter 5—*Exploratory Analyses of Safety Data* can be performed on dependent and independent variables to identify the outliers before developing the model (note that, in a technical sense, these are called high leverage observations instead of outliers). These observations should be investigated and either corrected or excluded from the data. During modeling, several measures can be used to identify the outliers. For example, Srinivasan and Bauer (2013) proposed using Cook's Distance statistic, a measure that is based on the comparison of the change in fitted values with and without that observation. As a rule of thumb, an outlier is a point that has a Cook's Distance greater than 4/n (where n is the total number of observations in the data). CURE plots can also be used to identify outliers. A vertical jump in the CURE plot indicates the presence of an outlier (Hauer, 2004). Another simple but effective technique is to use the standardized residuals. The standardized residuals are calculated by dividing the residuals with the variance of predicted crashes. Then, these standardized residuals are plotted against the predicted crashes to identify extreme observations.

It is important to note that outliers that are identified as such should be observations that are clearly identified as being erroneous. By erroneous, we mean that some of the variables could have included a coding error or an observation that is mistakenly included in a group of sites with similar characteristics. For example, for the former, the traffic flow on an urban segment shows a value of 4000 vehicles per day, while all the other adjacent segments have 40,000 vehicles per day on average or, for the latter, a modern roundabout that is incorrectly included among signalized intersections. Those may need to be either removed or further investigated. As crashes and other related data are probabilistic in nature, it is expected that some observations will be classified as outliers, say beyond the 95% CIs. Hence, as a general rule, if 5% of the observations are located beyond the 95% CIs, this would be considered acceptable. However, if 10% of the observations are situated beyond the 95% CIs, then there may be some problems with the data and/or the model. Furthermore, removing outliers may lead to additionally identified outliers when the model is reestimated with the reduced dataset. This is why removing outliers should be done on a case-by-case basis and for reasons described earlier. Another general rule states that observations located beyond three standard deviations away from the mean of the crash-frequency model or the sample mean should be thoroughly investigated, as those are most likely to be true outliers.

6.5.6 Model transferability

When a crash-frequency or crash-severity model developed in one jurisdiction is desired to be transferred to use in another jurisdiction, an adjustment factor needs to be developed (the HSM calls the process as calibration and the adjustment factor as a calibration factor). As crash frequency and its dispersion vary substantially from one jurisdiction to next, it is essential to calibrate crash-frequency models when they are applied to a new jurisdiction. In other words, calibration is a tool to account for the differences in factors that are not considered or cannot be considered in the development of crash-frequency models, such as weather, driver behavior, and reportability criteria between jurisdictions.

The following equation is used for calculating the calibration factor (AASHTO, 2010).

$$C = \frac{\sum_{all\ sites} y}{\sum_{all\ sites} \mu} \tag{6.18}$$

where C is the calibration factor, y is the number of observed crashes, and μ is the number of unadjusted predicted crashes from the crash-frequency model.

By estimating a calibration factor in the above manner, a straight line relation is assumed between observed crashes and predicted crashes. Hauer (2013) suggests that the safety analyst should examine whether or not this relationship is indeed a straight line by estimating the following model:

$$y = C \times \mu^{\beta} \tag{6.19}$$

where β is the parameter to be estimated using a regression analysis. If the relationship between observed crashes and predicted crashes is a straight line, then β will be close to 1.0.

Annual calibration factors for several years will enable the assessment of whether or not there is a time trend leading to a calibration function (i.e., a function of time) for use in the past and future analyses. Developing the temporal calibration factors (i.e., recalibrating the predictive models over time) is inevitable as the characteristics of crash data are likely to change over time. Although more efficient than fitting a new model, recalibrating the predictive models can still be a challenging task. Furthermore, the effort put into recalibrating models could be wasted if a recalibration is not required at that point of time. Consequently, it is important to know when or how often CFMs should be recalibrated to avoid unnecessary calibration efforts or expenses. Lord et al. (2016) provided a procedure to determine when the analyst is advised to recalibrate the predictive models. The procedure is based on general characteristics of data that will be used for recalibration of the predictive models. The procedure recommends using the network-level traffic flow and mileage for developing an approximate calibration factor (called as C-proxy) with which the authors showed when to recalibrate the predictive models. For more details about the procedure, the reader is referred to Lord et al. (2016).

Similar to crash frequency models, calibration is also needed for crash-severity models (also called severity distribution functions in the HSM) developed in one jurisdiction to be used in another jurisdiction. The crash-severity model usually includes explanatory variables, such as the site geometric design features, traffic control elements, and traffic flow characteristics to estimate the likelihood of each severity outcome. As the crash-severity model accounts for all severity outcomes together, a single change in a variable could result in simply shifting the number of crashes between different severity alternatives. The following equation is used for calculating the calibration factor for crash-severity models (Bonneson et al., 2012).

$$C = \frac{\sum_{all\ sites} y_s / \sum_{all\ sites} \mu_s}{\sum_{all\ sites} (y - y_s) / \sum_{all\ sites} (\mu - \mu_s)} \tag{6.20}$$

where C is the calibration factor; y_s is the number of observed crashes with severity s; y is the total number of observed crashes; and, μ_s is the number of unadjusted predicted crashes with severity s, and μ is the total unadjusted predicted crashes.

Exercise 6.3

Use the HSM crash frequency model and the dataset in Exercise 6.1, develop the adjustment factor for model transferability.

First, select the CFM from the HSM for predicting the crashes on rural multilane undivided highway segments (AASHTO, 2010).

$$\mu = \exp(-9.653 + 1.176 \times \ln(AADT) + \ln(L))$$

Second, determine if the sites meet the baseline conditions in the HSM. As the sideslope and lighting information are not available, for simplicity, let us assume that they meet the baseline conditions (note that the analyst must collect the information about these variables before developing the adjustment factors). After considering the sites with 12-foot lanes, there are too few sites that meet the HSM baseline condition for shoulder width of 6 ft. Instead, we can consider the sites that have a shoulder width of 8 ft.

Third, apply the CMF for the shoulder width and calculate the predicted crashes. As the CFM estimates annual crashes, we must multiply with the number of years of which crash data were observed. In this case, it will be 5 years.

Fourth, using Eq. (6.18), estimate the adjustment factor.

$$C = \frac{\sum_{all\ sites} y}{\sum_{all\ sites} \mu} = \frac{876}{697} = 1.26$$

If the sideslope variable CMF was considered, then the predicted crashes would be higher and the adjustment factor moves closer to 1.0.

6.6 Other study types

This section presents three other types of cross-sectional studies that have been used in crash data analysis.

6.6.1 Cohort studies

Cohort studies are one type of longitudinal studies in which cohorts (e.g., a group of people who received driver training and another group without the training) are first identified and followed at intervals through time until the outcome of interest (e.g., driver injury) occurs. Cohort studies have the potential to provide the strongest scientific evidence to assess causality (Song and Chung, 2010). These studies are mainly used to analyze rare exposures and allow examining multiple outcomes simultaneously. For instance, using the data collected in New Zealand, Whitlock et al. (2003) first investigated the relationship between motor vehicle driver injury and socioeconomic status. Using the same data, the authors also investigated the relationship between motor vehicle driver injury and marital status (Whitlock et al., 2004). The main disadvantages of cohort studies include the need for a large sample size and long follow-up periods, which increase the risk of subjects to drop out of the study. For example, Whitlock et al. (2003) and Whitlock et al. (2004) used data collected from 10,525 drivers during the period 1988—98.

Two important types of cohort studies are:

1. **Prospective Cohort Study**: In prospective cohort studies, the exposure data from the recruited subjects are first collected before the development of the outcomes of interest. They are followed in future times until the occurrence of the outcome of interest.
2. **Retrospective Cohort Study**: In retrospective cohort studies, the investigation starts with the study of the subject's past to identify the exposure after some people have already experienced the outcome of interest. These studies may have multiple exposures, such as the example described earlier.

6.6.2 Case-control studies

Case-control studies are sometimes used as alternatives to before-after studies because the latter requires a large amount of data and may not isolate the safety effect of a single intervention when other adjustments are implemented at the same time (i.e., the safety effect may not necessarily due to the intervention but may be due to other confounding factors). Case-control studies are also extensively applied in real-time crash prediction studies to control confounding variables and unobserved factors associated with crash outcome, see Chapter 10 - *Capacity, Mobility and Safety* (Section 10.6) for more details. In case-control studies, observations or participants are selected based on the outcome they experienced during a selected study period. For example, roadway segments that experienced a particular type of crashes are selected as cases, whereas others that have not experienced any such crash types are selected as controls to study the effect of one or more risk factors. The case-control method is particularly useful for analyzing rare outcomes such as traffic crashes. The characteristics of controls are often matched to those of cases to ensure that the cases and controls are similar, however, nonmatched controls are also used. Matching of controls can be done in two ways (Pokorny et al., 2020): (1) one-to-one matching where one control or a specific number of controls are matched to one case, and (2) frequency-matching where matching of controls is based upon the distributions of the characteristics among the cases. Woodward (2013) recommended that a maximum of four controls per one case can be used as adding more controls does not increase the power of the study. The outcomes in case and control groups are presented in the form of two-way contingency tables and the odds ratio is calculated to interpret the results. A detailed description of the odds ratio is presented in Chapter 5—*Exploratory Analyses of Safety Data*. Various studies have used the case-control method in evaluating the safety effectiveness of treatments. Bakiri et al. (2013) evaluated the risk associated with driver distraction on the traffic crash injury by interviewing the injured drivers at the emergency room. Kuypers et al. (2012) and Lacey et al. (2016) used the case-control study method to assess the risk of involving in a traffic crash after using alcohol, drugs, or both. Kuypers et al. (2012) also investigated the data further to determine the concentrations at which this risk is significantly increased.

6.6.3 Randomized control trials

A randomized controlled trial, also called randomized control trial (RCT), is a prospective, comparative, and quantitative study/experiment, which aims to reduce certain sources of bias when testing the effectiveness of interventions. The RCT is performed under controlled conditions with random allocation of interventions to comparison groups. The first group or experimental group has the intervention being assessed, whereas the second group or control group has an alternative condition, such as a placebo or no intervention. The groups are monitored and assessed under the experiment design conditions to see how effective the experimental intervention is. The effectiveness of the treatment is assessed in comparison to the control. It is possible to include more than one treatment group or more than one control group.

The RCT is the most rigorous and robust research method for determining whether or not a cause–effect relation exists between an intervention and an outcome. High-quality evidence can be generated by performing a randomized controlled trial when evaluating the effectiveness of a safety intervention.

For estimating the effect of a variable on crash risk, the ideal method would be to conduct an experimental study on an existing road network (Hauer, 2010). However, in highway safety research, it is unethical and uneconomical to conduct an experiment in a real traffic environment (Gross, 2013). The RCT is generally adopted to study the behavioral measures of the drivers (see Ker et al. (2005)).

References

AASHTO, 2010. Highway Safety Manual, first ed. American Association of State Highway and Transportation Officials, Washington, D.C.

Aptel, I., Salmi, L.R., Masson, F., Bourdé, A., Henrion, G., Erny, P., 1999. Road accident statistics: discrepancies between police and hospital data in a French island. Accid. Anal. Prev. 31 (1—2), 101—108.

Ash, J.E., Zou, Y., Lord, D., Wang, Y., 2020. Comparison of confidence and prediction intervals for different mixed-poisson regression models. J. Transport. Saf. Secur. 1—23.

Bahadorimonfared, A., Soori, H., Mehrabi, Y., Delpisheh, A., Esmaili, A., Salehi, M., Bakhtiyari, M., 2013. Trends of fatal road traffic injuries in Iran (2004—2011). PLoS One 8 (5), e65198.

Bakiri, S., Galéra, C., Lagarde, E., Laborey, M., Contrand, B., Ribéreau-Gayon, R., Salmi, L.-R., Gabaude, C., Fort, A., Maury, B., 2013. Distraction and driving: results from a case—control responsibility study of traffic crash injured drivers interviewed at the emergency room. Accid. Anal. Prev. 59, 588—592.

Bauer, K., Harwood, D., 2013. Safety effects of horizontal curve and grade combinations on rural two-lane highways. Transport. Res. Rec. J. Transport. Res. Board (2398) 37—49.

Blincoe, L.J., Seay, A.G., Zaloshnja, E., Miller, T.R., Romano, E.O., Luchter, S., Spicer, R.S., 2002. The Economic Impact of Motor Vehicle Crashes, 2000. National Highway Traffic Safety Administration, United States.

Bonneson, J., Geedipally, S., Pratt, M., Lord, D., 2012. Safety Prediction Methodology and Analysis Tool for Freeways and Interchanges. Texas Transportation Institute, College Station, Texas, pp. 17—45.

Brijs, T., Karlis, D., Wets, G., 2008. Studying the effect of weather conditions on daily crash counts using a discrete time-series model. Accid. Anal. Prev. 40 (3), 1180—1190.

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

Carson, J., Mannering, F., 2001. The effect of ice warning signs on ice-accident frequencies and severities. Accid. Anal. Prev. 33 (1), 99—109.

Davis, G.A., 2004. Possible aggregation biases in road safety research and a mechanism approach to accident modeling. Accid. Anal. Prev. 36 (6), 1119—1127.

Diop, M.L., Diop, A., Kâ, A., 2018. A negative binomial mixture integer-valued garch model. Afrika Statistika 13 (2), 1645—1666.

Elvik, R., Mysen, A., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. Transport. Res. Rec. 1665 (1), 133—140.

Eze, C., Okonkwo, C., 2018. On the modelling of road traffic crashes: a case of sarima models. J. Adv. Res. Math. Stat. (ISSN: 2208-2409) 5 (8), 15—35.

Geedipally, S.R., Lord, D., Park, B.-J., 2009. Analyzing different parameterizations of the varying dispersion parameter as a function of segment length. Transport. Res. Rec. 2103 (1), 108—118.

Ghédira, A., Kammoun, K., Saad, C.B., 2018. Temporal analysis of road accidents by arima model: case of Tunisia. Int. J. Innovat. Appl. Stud. 24 (4), 1544—1553.

Gross, F., 2013. Case-control analysis in highway safety: accounting for sites with multiple crashes. Accid. Anal. Prev. 61, 87—96.

Harwood, D., Council, F.M., Hauer, E., Hughes, W.E., Vogt, A., 2000. Prediction of the Expected Safety Performance of Rural Two-Lane Highways. Midwest Research Institute, Kansas City, Missouri.

Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in empirical bayes estimation. Accid. Anal. Prev. 33 (6), 799—808.

Hauer, E., 2004. Statistical road safety modeling. Transport. Res. Rec. J. Transport. Res. Board 1897, 81—87.

Hauer, E., 2006. The frequency-severity indeterminacy. Accid. Anal. Prev. 38 (1), 78−83.

Hauer, E., 2010. Cause, effect and regression in road safety: a case study. Accid. Anal. Prev. 42 (4), 1128−1135.

Hauer, E. (2013), Safety Performance Functions: A Workshop, Baton Rouge, Louisiana, July 16−18, 2013.

Hauer, E., 2015. The Art of Regression Modeling in Road Safety. Springer, USA.

Hauer, E., Hakkert, A., 1988. Extent and some implications of incomplete accident reporting. Transport. Res. Rec. 1185 (1−10), 17.

Hausman, J.A., 1978. Specification tests in econometrics. Econometrica: J. Econometric Soc. 1251−1271.

Hausman, J.A., Hall, B.H., Griliches, Z., 1984. Econometric Models for Count Data with an Application to the Patents-R&D Relationship. National Bureau of Economic Research.

Houston, D.J., Richardson Jr., L.E., 2002. Traffic safety and the switch to a primary seat belt law: the California experience. Accid. Anal. Prev. 34 (6), 743−751.

Ihueze, C.C., Onwurah, U.O., 2018. Road traffic accidents prediction modelling: an analysis of anambra state, Nigeria. Accid. Anal. Prev. 112, 21−29.

Jaccard, J., Turrisi, R., 2003. Interaction Effects in Multiple Regression. Sage.

Ker, K., Roberts, I., Collier, T., Beyer, F., Bunn, F., Frost, C., 2005. Post-licence driver education for the prevention of road traffic crashes: a systematic review of randomised controlled trials. Accid. Anal. Prev. 37 (2), 305−313.

Kim, D.-G., Washington, S., 2006. The significance of endogeneity problems in crash models: an examination of left-turn lanes in intersection crash models. Accid. Anal. Prev. 38 (6), 1094−1100.

Kumara, S., Chin, H.C., 2005. Application of Poisson underreporting model to examine crash frequencies at signalized three-legged intersections. Transport. Res. Rec. 1908 (1), 46−50.

Kuypers, K.P.C., Legrand, S.-A., Ramaekers, J.G., Verstraete, A.G., 2012. A case-control study estimating accident risk for alcohol, medicines and illegal drugs. PLoS One 7 (8), e43496.

Lacey, J.H., Kelley-Baker, T., Berning, A., Romano, E., Ramirez, A., Yao, J., Moore, C., Brainard, K., Carr, K., Pell, K., 2016. Drug and Alcohol Crash Risk: A Case-Control Study. National Highway Traffic Safety Administration. Office of …, United States.

Lee, J., Abdel-Aty, A., Park, J., 2018. Investigation of associations between marijuana law changes and marijuana-involved fatal traffic crashes: a state-level analysis. J. Transport Health 10, 194−202.

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., 2008. Methodology for estimating the variance and confidence intervals for the estimate of the product of baseline models and amfs. Accid. Anal. Prev. 40 (3), 1013−1017.

Lord, D., Geedipally, S.R., 2018. Safety Prediction with Datasets Characterised with Excess Zero Responses and Long Tails. Safe Mobility: Challenges, Methodology and Solutions. Emerald Publishing Limited.

Lord, D., Geedipally, S.R., Guikema, S.D., 2010. Extension of the application of conway-maxwell-poisson models: analyzing traffic crash data exhibiting underdispersion. Risk Anal. 30 (8), 1268−1276.

Lord, D., Geedipally, S.R., Shirazi, M., Center, A., 2016. Improved Guidelines for Estimating the Highway Safety Manual Calibration Factors. University Transportation Centers Program (US).

Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the conway-maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. Accid. Anal. Prev. 40 (3), 1123−1134.

Lord, D., Kuo, P.-F., 2012. Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures. Accid. Anal. Prev. 47, 52−63.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transport. 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., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transport. Res. Rec. 1717 (1), 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.

Ma, P., 2009. Bayesian Analysis of Underreporting Poisson Regression Model with an Application to Traffic Crashes on Two-Lane Highways.

Makridakis, S., Wheelwright, S.C., Hyndman, R.J., 2008. Forecasting Methods and Applications. John Wiley & Sons.

Mannering, F., 2018. Cross-Sectional Modelling. Safe Mobility: Challenges, Methodology and Solutions, p. 257.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash flow relationships for intersections - dispersion parameter, functional form, and bayes versus empirical bayes methods. Transport. Res. Rec. J. Transport. Res. Board 1840, 31−40.

Miranda-Moreno, L.F., Lord, D., Fu, L., 2008. Evaluation of Alternative Hyper-Priors for Bayesian Road Safety Analysis.

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accid. Anal. Prev. 39 (3), 459−468.

Park, B.-J., Lord, D., 2008. Adjustment for maximum likelihood estimate of negative binomial dispersion parameter. Transport. Res. Rec. 2061 (1), 9−19.

Patil, S., Geedipally, S.R., Lord, D., 2012. Analysis of crash severities using nested logit model—accounting for the underreporting of crashes. Accid. Anal. Prev. 45, 646−653.

Pokorny, P., Jensen, J.K., Gross, F., Pitera, K., 2020. Safety effects of traffic lane and shoulder widths on two-lane undivided rural roads: a matched case-control study from Norway. Accid. Anal. Prev. 144, 105614.

Pratt, M.P., Geedipally, S.R., Wilson, B., Das, S., Brewer, M., Lord, D., 2018. Pavement Safety-Based Guidelines for Horizontal Curve Safety.

Quddus, M., 2018. Time-series regression models for analysing transport safety data. In: Safe Mobility: Challenges, Methodology and Solutions, p. 279.

Quddus, M.A., 2008. Time series count data models: an empirical application to traffic accidents. Accid. Anal. Prev. 40 (5), 1732−1741.

Quddus, M.A., 2016. Non-Gaussian Interrupted Time Series Regression Analysis for Evaluating the Effect of Smart Motorways on Road Traffic Accidents.

Rock, S.M., 1996. Impact of the Illinois child passenger protection act: a retrospective look. Accid. Anal. Prev. 28 (4), 487−492.

Rosenbaum, P.R., 2010. Design of Observational Studies. Springer.

Rosenbaum, P.R., 2017. Observation and Experiment. Harvard University Press.

Saha, K., Paul, S., 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. Biometrics 61 (1), 179−185.

Shirazi, M., Geedipally, S.R., Lord, D., 2020. A Simulation Analysis to Study the Temporal and Spatial Aggregations of Safety Datasets with Excess Zero Observations.

Song, J.W., Chung, K.C., 2010. Observational studies: cohort and case-control studies. Plast. Reconstr. Surg. 126 (6), 2234.

Srinivasan, R., Bauer, K., 2013. Safety Performance Function Development Guide: Developing Jurisdiction-Specific spfs. FHWA, Washington DC.

Usman, T., Fu, L., Miranda-Moreno, L.F., 2011. Accident prediction models for winter road safety: does temporal aggregation of data matter? Transport. Res. Rec. 2237 (1), 144−151.

Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. Statistical and Econometric Methods for Transportation Data Analysis. CRC Press.

Washington, S., Karlaftis, M.G., Mannering, F.L., 2011. Statistical and Econometric Methods for Transportation Data Analysis Boca Raton, FL. CRC Press, Boca Raton, FL.

Whitlock, G., Norton, R., Clark, T., Jackson, R., Macmahon, S., 2004. Motor vehicle driver injury and marital status: a cohort study with prospective and retrospective driver injuries. Inj. Prev. 10 (1), 33−36.

Whitlock, G., Norton, R., Clark, T., Pledger, M., Jackson, R., Macmahon, S., 2003. Motor vehicle driver injury and socioeconomic status: a cohort study with prospective and retrospective driver injuries. J. Epidemiol. Community Health 57 (7), 512−516.

Wood, G., 2002. Generalised linear accident models and goodness of fit testing. Accid. Anal. Prev. 34 (4), 417−427.

Wood, G., 2005. Confidence and prediction intervals for generalised linear accident models. Accid. Anal. Prev. 37 (2), 267−273.

Woodward, M., 2013. Epidemiology: Study Design and Data Analysis. CRC Press.

Wu, L., Lord, D., 2017. Examining the influence of link function misspecification in conventional regression models for developing crash modification factors. Accid. Anal. Prev. 102, 123−135.

Wu, L., Lord, D., Zou, Y., 2015. Validation of crash modification factors derived from cross-sectional studies with regression models. Transport. Res. Rec. J. Transport. Res. Board 2514, 88−96.

Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accid. Anal. Prev. 40 (4), 1320−1329.

Yannis, G., Karlaftis, M.G., Year. Weather effects on daily traffic accidents and fatalities: a time series count data approach. In: Proceedings of the Proceedings of the 89th Annual Meeting of the Transportation Research Board, pp. 14.

Ye, F., Garcia, T.P., Pourahmadi, M., Lord, D., 2012. Extension of negative binomial garch model: analyzing effects of gasoline price and miles traveled on fatal crashes involving intoxicated drivers in Texas. Transport. Res. Rec. 2279 (1), 31−39.

Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit. Transport. Res. Rec. 2241 (1), 51−58.

Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit and mixed logit models. Anal. Methods Accid. Res. 1 (0), 72−85.

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

Zhang, X., Pang, Y., Cui, M., Stallones, L., Xiang, H., 2015. Forecasting mortality of road traffic injuries in China using seasonal autoregressive integrated moving average model. Ann. Epidemiol. 25 (2), 101−106.

Zhu, F., 2011. A negative binomial integer-valued garch model. J. Time Anal. 32 (1), 54−67.

Zou, Y., Geedipally, S.R., Lord, D., 2013. Evaluating the double Poisson generalized linear model. Accid. Anal. Prev. 59 (0), 497−505.