Cross-Sectional Studies

Fall 2021

Overview

- 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).
- 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.

Types of data - 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.
- 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.

Types of data - Time-series data

- **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.
- **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.
- **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.

Types of data - Time-series data

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

The autoregressive integrated moving average (ARIMA) model is used for nonstationary time series to achieve stationarity by differencing the time series one or more times.

Negative Binomial Integer-valued Generalized Autoregressive Conditional Heteroscedastic (NBINGARCH) model should be used when overdispersion in time-series count data is observed.

Generalized linear autoregressive and moving average (GLARMA) is another model that was applied for modeling time series of crash counts with covariates.

Types of data - 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.
- 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.

Types of data - Cross-sectional data

- Primarily, there are three advantages of cross-sectional data:
  - They provide a more robust predictive model than panel data 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.

Types of data - Cross-sectional data

- 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, 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 the same point in space.
- 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.

Types of data - Cross-sectional 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 is the accuracy of average annual daily traffic (AADT) in most highway safety databases.
- It is common for a segment's AADT volume to be missing for one or more years. Hence, you can take the average over the study duration.

Types of data - Cross-sectional data

- 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.

Types of data - 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.

Types of data - Panel data

- 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).
  - 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).
- Panel data modeling is recommended in cases when the variables are observed over time.

Types of data - Panel data

- Panel data modeling has the following advantages:
  - 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.
  - 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.
- 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** (GEE) can be used for handling the correlation.

Data and modeling issues – Overdispersion and Underdispersion

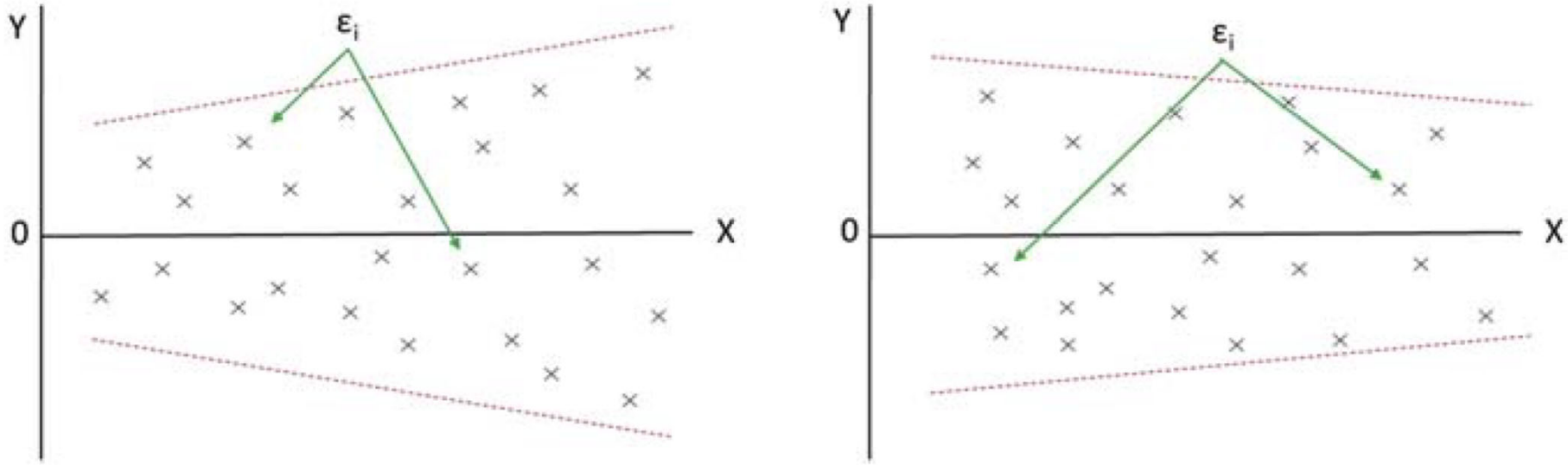


FIGURE 3.1 Overdispersed (left) and underdispersed (right) residuals.

Data and modeling issues – Small Sample Size and Low Sample Mean

- Due to the 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.
- Data characterized by small sample size and low sample-mean can cause estimation problems in traditional count (crash-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.
- 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.

Data and modeling issues - 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.
- 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.

Data and modeling issues – 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.
- 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

Data and modeling issues – Endogenous variables

- An endogenous variable is an explanatory variable whose value is determined or influenced by one or more variables in 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).

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. 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.

Data aggregation

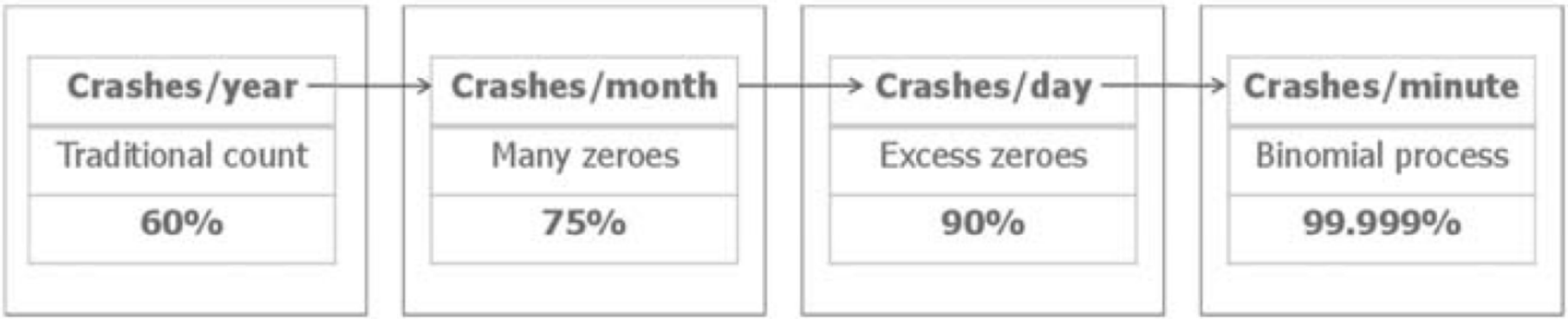


FIGURE 6.1 Percentage of zero responses when changing the time scale (Lord and Geedipally, 2018).

Data aggregation

- 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 observations, in which the traditional negative binomial (NB) model may not be appropriate for the safety analysis.
  - On the other hand, too much aggregation may result in loss of information, although it may make the NB model a better alternative (called "aggregation bias" or "ecological fallacy").

Data aggregation

- Recommended conservative criteria for balance:
  - 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.

Functional Forms

- 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 – discussed in Chapter 2 and in the next slide).
- 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.
- There is a body of work that suggests that nonlinear functions better characterize the relationships between crash frequencies and explanatory variables.

Functional Forms

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

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: $1/\left(1 + \beta_1 e^{\beta_2 X}\right) - 1/(1 + \beta_1)$	**7** Quadratic: $1 + \beta_1 X + \beta_2 X^2$
4 Weibull: $1 - e^{-(X/\beta_1)^{\beta_2}}$	

As discussed before, the selection of the estimation method, MLE or Bayes method, will be governed, in part, by the complexity of the functional form.

Functional Forms

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) with the variable of interest and the crashes (after accounting for exposure).

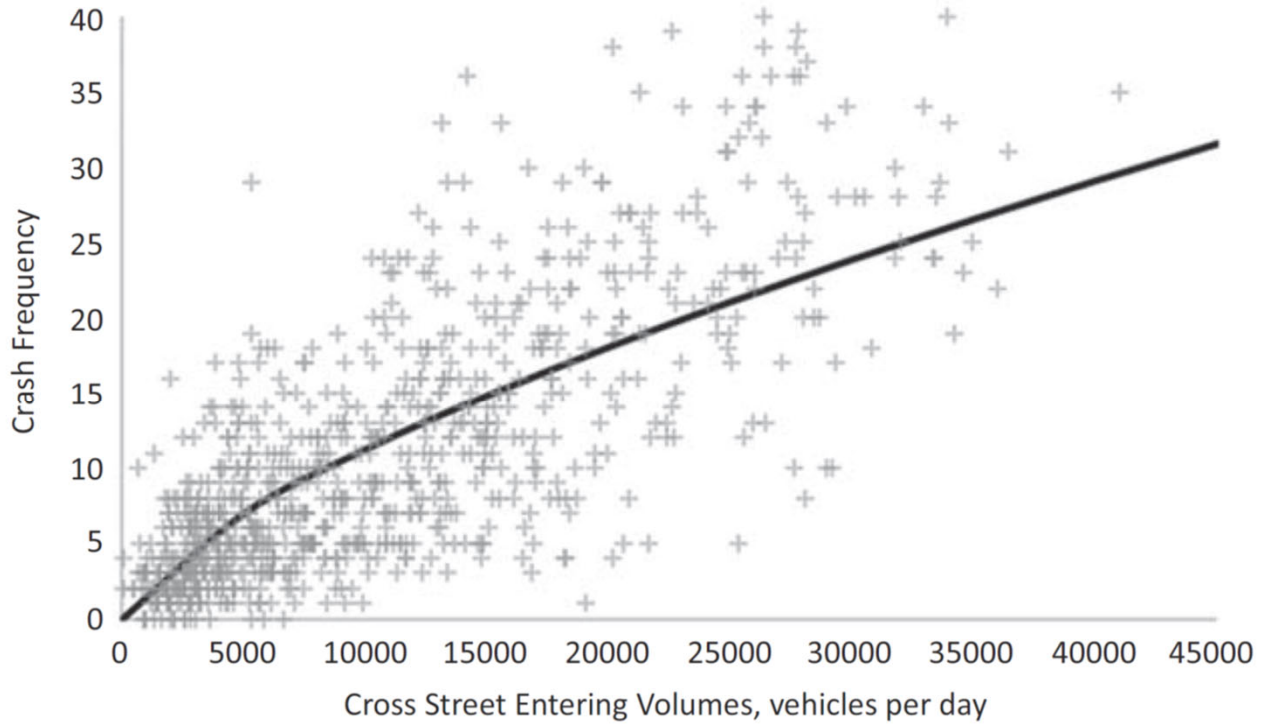


FIGURE 6.2 Relationship between cross street entering volumes and crash frequency.

Functional Forms - 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.

Functional Forms - Flow-only models

Road Segments

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

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

Intersections

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

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

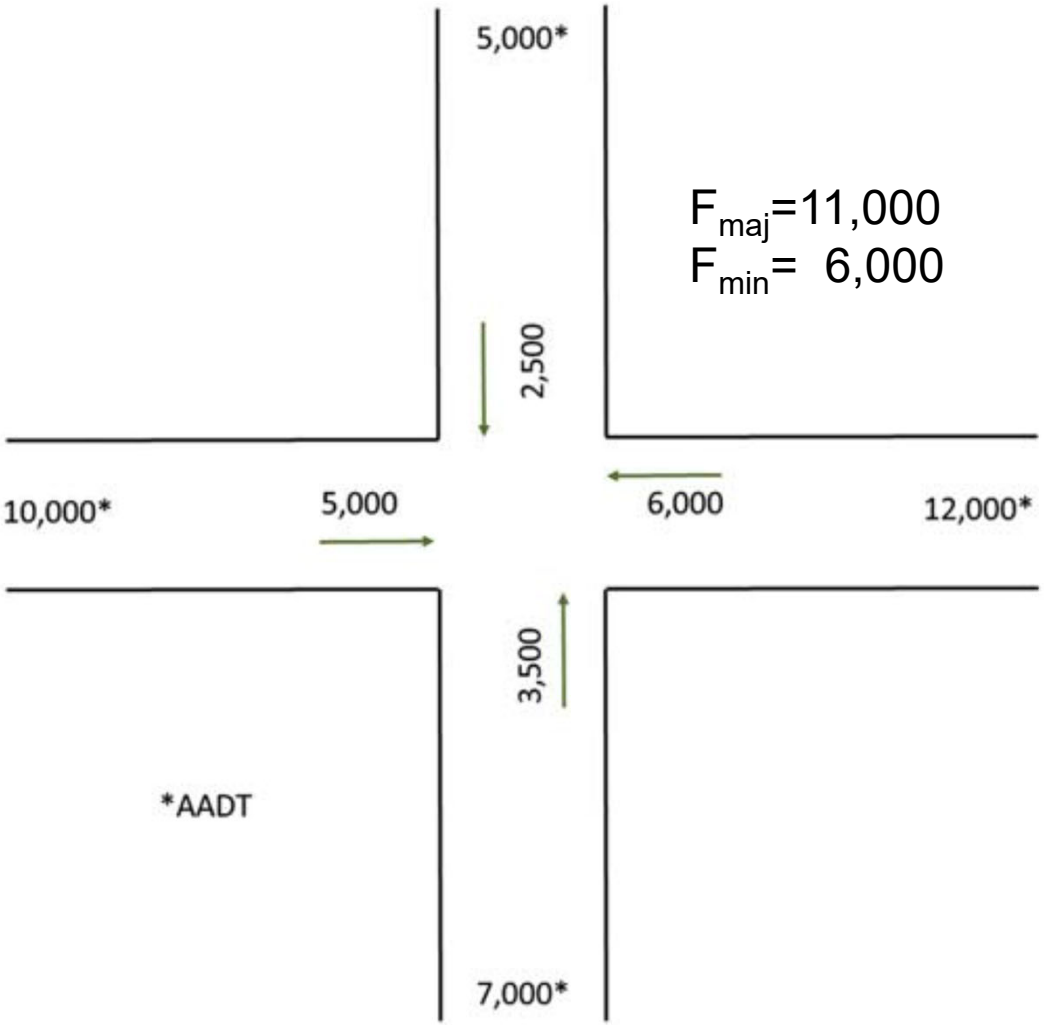
$$\mu_{int} = \beta_0 \times (AADT_{maj} + AADT_{min})^{\beta_1}$$

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

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

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

Functional Forms - Flow-only models



F_{maj}=11,000
F_{min}= 6,000

FIGURE 2.2 Entering flows in vehicles per day (AADT).

Functional Forms – 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 shown previously)

Functional Forms – Flow-only models with CMFs

$$\mu = \mu_b \times (CMF_1 \times CMF_2 \times \ldots . CMF_n)$$

where μ is the predicted crashes of an entity; μ_b is the predicted crashes for base conditions (note that the functional forms presented above 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.

This functional form is the one that has been adopted by the Highway Safety Manual (AASHTO) and the FHWA among others.

Functional Forms – 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.
- All the models described in Chapter 3 – Crash-Frequency Modeling and Chapter 4 - Crash-Severity Modeling would be applicable here.

Functional Forms – Model with covariates

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

where x_k is the safety-related variable (e.g., lane width, median width, turning lane) (k = 0, 1, 2 …). β_j are the regression coefficients (j = 0, 1, 2 ...).

Note that this functional form shows a simple log-linear relationship.

However, xi 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.

Variable selection

- This 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.
  - 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.
- Forward/Backward Selection (similar process)

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.
- 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), however, 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.
- Others since then have claimed that the variance function could be made dependent upon the variables in the model, not only segment length.

Crash variance and confidence intervals

Fixed: $$Var(y) = \mu + \alpha\mu^2$$

Function of segment length: $$\alpha_i = e^{\gamma_0} L_i^{\gamma_1}$$

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

Function of covariates: $$\alpha_i = e^{\gamma_0 + \gamma_1 \times AADT_{maj,i} + \gamma_2 \times AADT_{min,i} + \gamma_3 \times AADT_{min},i/AADT_{maj,i}}$$

$$\alpha_i = \exp\left(\mathbf{z}_i'\boldsymbol{\gamma} + \varpi_i\right)$$

Crash variance and confidence intervals

Parameter	Intervals
μ	$\left[\frac{\hat{\mu}}{e^{1.96\sqrt{Var(\hat{\eta})}}}, \hat{\mu}e^{1.96\sqrt{Var(\hat{\eta})}}\right]$
m	$\left[\max\left\{0, \hat{\mu} - 1.96\sqrt{\hat{\mu}^2\,\mathrm{var}(\hat{\eta}) + \frac{\hat{\mu}^2\,\mathrm{var}(\hat{\eta}) + \hat{\mu}^2}{\phi}}\right\}, \hat{\mu} + 1.96\sqrt{\hat{\mu}^2\,\mathrm{var}(\hat{\eta}) + \frac{\hat{\mu}^2\,\mathrm{var}(\hat{\eta}) + \hat{\mu}^2}{\phi}}\right]$
y	$\left[0, \left\lfloor \hat{\mu} + \sqrt{19}\sqrt{\hat{\mu}^2 Var(\hat{\eta}) + \frac{\hat{\mu}^2 Var(\hat{\eta}) + \hat{\mu}^2}{\phi}} + \hat{\mu} \right\rfloor\right]$
Note: $Var(\hat{\eta}) = XI^{-1}X^T$ 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	

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).

Sample Size

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

Sample Size

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

Other study types

- **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.
  - Prospective Cohort Study and Retrospective Cohort Study
- 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.
- 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.
  - Most rigorous analytic method, but in highway safety research, it is unethical and uneconomical to conduct an experiment in a real traffic environment.