交通大数据

计数回归

- 郭延永、徐铖铖、李豪杰
- guoyanyong@seu.edu.cn

Statistical Models For Crash Data

- The normal distribution played an important role in estimating the coefficients and inferences of probabilistic models. Unfortunately, there are many practical situations where the normal assumption is not valid. Count data, binary response (0 or 1) or other continuous variables with positive and high-skewed distribution cannot be modeled with a normally distributed errors.
- Count data consist of non-negative integer values and are encountered frequently in the modeling of transportation-related phenomena.

Statistical Models For Crash Data

- Examples of count data variables in transportation include the number of driver route changes per week, the number of trip departure changes per week, drivers' frequency of use of intelligent transportation systems (ITS) technologies over some time period, number of vehicles waiting in a queue, and the number of accidents observed on road segments per year.
- A common mistake is to model count data as continuous data by applying standard least squares regression. This is not correct because regression models yield predicted values that are non-integers and can also predict values that are negative, both of which are inconsistent with count data. These limitations make standard regression analysis inappropriate for modeling count data without modifying dependent variables.

Statistical Models For Crash Data

- The **generalized linear model** (GLM) was developed to allow fitting regression models for univariate response data that follows a very general distribution called exponential family. This family includes the normal, binomial, negative binomial, gamma, etc.
- 广义线性模型线性模型，通过联结函数建立响应变量的数学期望值与线性组合的预测变量之间的关系。其特点是不强行改变数据的自然度量，数据可以具有非线性和非恒定方差结构。是线性模型在研究响应值的非正态分布以及非线性模型简洁直接的线性转化时的一种发展。

Statistical Models For Crash Data

- **线性回归模型 (linear regression model)：**

$y_i = \beta_o + \beta_1 x_{1i} + \varepsilon_i$

- **特点：**
- 1. 响应变量y_i和误差项ε_i正态性：响应变量y_i和误差项ε_i服从正态分布，且ε_i具有零均值，同方差的特性。
- 2. 未知参数βi：运用最小二乘法或极大似然法解出的未知参数的估计值则具有正态性。
- 3.研究对象：线性回归模型主要研究响应变量的均值E[Y]。
- 4. 联接方式：

$$E[y_i]=f(\beta_o + \beta_1 x_{1i})= \beta_o + \beta_1 x_{1i}$$

Statistical Models For Crash Data

- **广义线性模型 (generalized linear model)：**
- **特点：**

1.响应变量的分布推广至指数分散族 (exponential dispersion family)：比如**泊松分布、二项分布、负二项分布、伽玛分布、逆高斯分布**。

3. 未知参数βi：未知参数βi认为是未知且不具有随机性的常数。

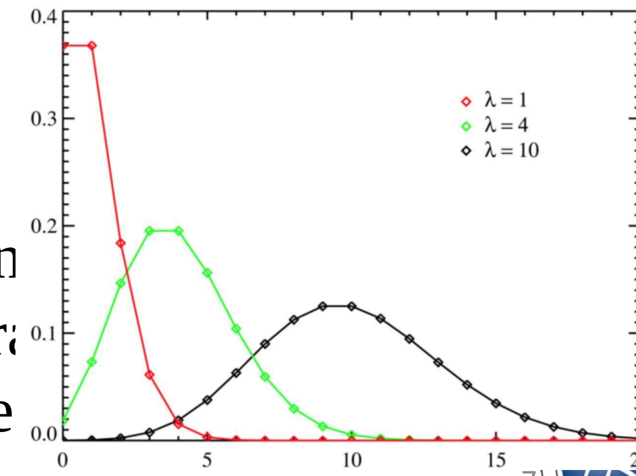
2.研究对象：广义线性模型的主要研究对象仍然是响应变量的均值E[Y]。

4.联接方式：广义线性模型里采用的联结函数 (link function) 理论上可以是任意的，而不再局限于f(x)=x。联结函数的选取必然地必须适应于具体的研究案例。

Poisson Regression Model

- Poisson regression is applied to a wide range of transportation count data. The Poisson distribution approximates rare-event count data, such as accident occurrence, failures in manufacturing, and number of vehicles waiting in a queue
- Consider the number of accidents occurring per year at various intersections in a city. In a Poisson regression model, the probability of intersection *i* having y_i accidents per year (where y_i is a non-negative integer) is given by:

$$P(y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$



where $P(y_i)$ is the probability of intersection accidents per year; and λ_i is the Poisson para intersection *i*, which is **equal to** the expecte accidents per year at intersection *i*, $E[y_i]$.

Poisson Regression Model

- Poisson regression models are estimated by specifying the Poisson parameter λ_i (the expected number of events per period) as a function of explanatory variables (For the intersection accident example, explanatory variables might include the geometric conditions of the intersections, signalization, pavement types, visibility, and so on.).

$$\lambda_i = EXP(\beta_0 + \beta X_i) \qquad u_i = \beta_0 + \beta X_i$$

- where $\boldsymbol{X_i}$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of estimable parameters.

Poisson Regression Model

- In this formulation, the expected number of events per period is given by $E[y_i]= \lambda_i=EXP(\boldsymbol{\beta}\mathbf{X}_i)$. This model is estimable by standard maximum likelihood methods, with the likelihood function given as

$$L(\beta) = \prod_i \frac{EXP[-EXP(\beta X_i)][EXP(\beta X_i)]^{y_i}}{y_i!}$$

- The log of the likelihood function is simpler to manipulate and more appropriate for estimation,

$$LL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[-EXP(\boldsymbol{\beta}\mathbf{X}_i) + y_i\boldsymbol{\beta}\mathbf{X}_i - LN(y_i!) \right]$$

- The estimated parameters are used to make inferences about the unknown population characteristics thought to impact the count process.

Poisson Regression Model

- To provide some insight into the implications of parameter estimation results, elasticities are computed to determine the marginal effects of the independent variables.
- Elasticities provide an estimate of the impact of a variable on the expected frequency and are interpreted as the effect of a 1% change in the variable on the expected frequency λ_i.
- For example, an elasticity of –1.32 is interpreted to mean that a 1% increase in the variable reduces the expected frequency by 1.32%.

Poisson Regression Model

- Elasticities are the correct way of evaluating the relative impact of each variable in the model. Elasticity of frequency λ_i is defined as

$$E_{x_{ik}}^{\lambda_i} = \frac{\partial \lambda_i}{\lambda_i} \times \frac{x_{ik}}{\partial x_{ik}} = \beta_k x_{ik}$$

  where E represents the elasticity, x_{ik} is the value of the kth independent variable for observation i, β_k is the estimated parameter for the kth independent variable and λ_i is the expected frequency for observation i.

- Note that elasticities are computed for each observation i. It is common to report a single elasticity as the average elasticity over all i.

Poisson Regression Model

- For indicator variables, a pseudo-elasticity is computed to estimate an approximate elasticity of the variables. The pseudo-elasticity gives the incremental change in frequency caused by changes in the indicator variables.
- The pseudo-elasticity for indicator variables, is computed as

$$E_{x_{ik}}^{\lambda_i} = \frac{EXP(\beta_k) - 1}{EXP(\beta_k)}$$

Poisson Regression Model

- When selecting among alternative models, GOF statistics should be considered along with model plausibility and agreement with expectations.
- There are numerous goodness-of-fit (GOF) statistics used to assess the fit of the Poisson regression model to observed data.
- The **likelihood ratio test** is a common test used to assess two competing models.
- It provides evidence in support of one model, usually a full or complete model, over another competing model that is restricted by having a reduced number of model parameters.

Poisson Regression Model

- The likelihood ratio test statistic is

$$X^2 = -2[LL(\beta_R) - LL(\beta_u)]$$

where $LL(\beta_R)$ is the log likelihood at convergence of the "restricted" model (sometimes considered to have all parameters in β equal to 0, or just to include the constant term, to test overall fit of the model), and $LL(\beta_U)$ is the log likelihood at convergence of the unrestricted model.

- The χ^2 statistic is χ^2 distributed with the degrees of freedom equal to the difference in the numbers of parameters in the restricted and unrestricted model (the difference in the number of parameters in the β_R and the β_U parameter vectors.

Poisson Regression Model

- The sum of model deviances, G^2, is equal to zero for a model with perfect fit. Note, however, that because observed y_i is an integer while the predicted expected value is continuous, a G^2 equal to zero is a theoretical lower bound.
- This statistic is given as

$$G^2 = 2\sum_{i=1}^{n} y_i LN\left(\frac{y_i}{\hat{\lambda}_i}\right)$$

Poisson Regression Model

- An equivalent measure to R^2 in ordinary least squares linear regression is not available for a Poisson regression model due to the nonlinearity of the conditional mean (E[y|X]) and heteroscedasticity in the regression.
- A similar statistic is based on standardized residuals,

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n} \left[\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}} \right]^2}{\sum_{i=1}^{n} \left[\frac{y_i - \bar{y}}{\sqrt{\bar{y}}} \right]^2}$$

where the numerator is similar to a sum of square errors and the denominator is similar to a total sum of squares.

Poisson Regression Model

- Another measure of overall model fit is the ρ^2 statistic. The ρ^2 statistic is

$$\rho^2 = 1 - \frac{LL(\beta)}{LL(0)}$$

where LL(β) is the log likelihood at convergence with parameter vector β and LL(0) is the initial log likelihood (with all parameters set to zero).

- The perfect model would have a likelihood function equal to one (all selected alternative outcomes would be predicted by the model with probability one, and the product of these across the observations would also be one) and the log likelihood would be zero, giving a ρ^2 of one.
- Thus the ρ^2 statistic is between zero and one and the closer it is to one, the more variance the estimated model is explaining.

Case study

Accident data from California (1993 to 1998) and Michigan (1993 to 1997) were collected (Vogt and Bared, 1998; Vogt, 1999). The data represent a culled data set from the original studies, which included data from four states across numerous time periods and over five different intersection types. A reduced set of explanatory variables is used for injury accidents on three-legged stop-controlled intersections with two lanes on the minor and four lanes on the major road. The accident data are thought to be approximately Poisson or negative binomial distributed, as suggested by previous studies on the subject (Miaou and Lum, 1993; Miaou 1994; Shankar et al., 1995; Poch and Mannering, 1996; Milton and Mannering, 1998; and Harwood et al., 2000). The variables in the study are summarized in Table 10.1.

Case study

Summary of Variables in California and Michigan Accident Data

Variable Abbreviation	Variable Description	Maximum/ Minimum Values	Mean of Observations	Standard Deviation of Observations
STATE	Indicator variable for state: 0 = California; 1 = Michigan	1/0	0.29	0.45
ACCIDENT	Count of injury accidents over observation period	13/0	2.62	3.36
AADT1	Average annual daily traffic on major road	33058/2367	12870	6798
AADT2	Average annual daily traffic on minor road	3001/15	596	679
MEDIAN	Median width on major road in feet	36/0	3.74	6.06
DRIVE	Number of driveways within 250 ft of intersection center	15/0	3.10	3.90

Case study

Poisson Regression of Injury Accident Data

Independent Variable	Estimated Parameter	t Statistic
Constant	–0.826	–3.57
Average annual daily traffic on major road	0.0000812	6.90
Average annual daily traffic on minor road	0.000550	7.38
Median width in feet	– 0.0600	– 2.73
Number of driveways within 250 ft of intersection	0.0748	4.54
Number of observations	84	
Restricted log likelihood (constant term only)	–246.18	
Log likelihood at convergence	–169.25	
Chi-squared (and associated p-value)	153.85 (<0.0000001)	
R_p-squared	0.4792	
G^2	176.5	

Case study

- A likelihood ratio test comparing the fitted model and the reduced model results in $X^2 = 153.85$, which is sufficient to reject the fit of the reduced model. Thus it is very unlikely (p-value less than 0.0000001) that randomness alone would produce the observed decrease in the log likelihood function.
- Note that G^2 is 186.48, which is only relevant in comparison to other competing models. The R^2_p is 0.48, which again serves as a comparison to competing models. **A model with higher X^2, lower G^2, and a higher R^2_p is sought in addition to a more appealing** set of individual predictor variables, model specification, and agreement with expectation and theory.

Case study

Average Elasticities of the Poisson Regression Model Shown in

Independent Variable	Elasticity
Average annual daily traffic on major road	1.045
Average annual daily traffic on minor road	0.327
Median width in feet	–0.228
Number of driveways within 250 ft of intersection	0.232

AADT1 > AADT2 > DRIVE >MEDIAN

Case study

$$E[y_i] = \lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i)$$

$$= EXP\begin{pmatrix} -0.83 + 0.00008(AADT1_i) \\ +0.0005(AADT2_i) - 0.06(MEDIAN_i) + 0.07(DRIVE_i) \end{pmatrix}.$$

(1) AADT1 = 33058; (2) AADT2 =3001; (3) MEDIAN = 30; (4) DRIVE = 3

$$E[y] = \lambda = \exp\begin{pmatrix} -0.83 + 0.00008 \times 33058 + 0.0005 \times 3001 \\ -0.06 \times 30 + 0.07 \times 3 \end{pmatrix}$$

$$= 5.613$$

Negative Binomial Model

- A common analysis error is a result of failing to satisfy the property of the Poisson distribution that restricts the mean and variance to be equal, when $E[y_i]=VAR[y_i]$.
- If this equality does not hold, the data are said to be under dispersed ($E[y_i]>VAR[y_i]$) or overdispersed ($E[y_i]<VAR[y_i]$), and the parameter vector is biased if corrective measures are not taken.
- Overdispersion can arise for a variety of reasons, depending on the phenomenon under investigation
- The primary reason in many studies is that variables influencing the Poisson rate across observations have been omitted from the regression.

Negative Binomial Model

- With the negative binomial model, the relationship between the observed crash count and explanatory variables is given as,

$$\lambda_i = EXP\left(\beta x_i + \varepsilon_i\right) = \exp(\boldsymbol{\beta X_i})\exp(\varepsilon_i)$$

where $EXP(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α^2.

- The addition of this term allows the variance to differ from the mean as below:

$$VAR[y_i] = E[y_i]\left[1+\alpha E[y_i]\right] = E[y_i] + \alpha E[y_i]^2$$

Negative Binomial Model

- The Poisson regression model is regarded as a limiting model of the negative binomial regression model as α approaches zero, which means that the selection between these two models is dependent on the value of α.
- The parameter α is often referred to as the overdispersion parameter.
- The negative binomial distribution has the form:

$$P(y_i)=\frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}\left(\frac{1}{1+\alpha\lambda_i}\right)^{\alpha^{-1}}$$

Negative Binomial Model

- The likelihood function of the negative binomial distribution is given by:

$$L(\lambda_i)=\prod_i \frac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha\lambda_i}{1+\alpha\lambda_i}\right)^{y_i}\left(\frac{1}{1+\alpha\lambda_i}\right)^{\alpha^{-1}}$$

- When the data are overdispersed, the estimated variance term is larger than under a true Poisson process. As overdispersion becomes larger, so does the estimated variance, and consequently all of the standard errors of parameter estimates become inflated.

Negative Binomial Model

Negative Binomial Regression of Injury Accident Data

Independent Variable	Estimated Parameter	t Statistic
Constant	–0.931	–2.37
Average annual daily traffic on major road	0.0000900	3.47
Average annual daily traffic on minor road	0.000610	3.09
Median width in feet	– 0.0670	–1.99
Number of driveways within 250 ft of intersection	0.0632	2.24
Overdispersion parameter, α	0.516	3.09
Number of observations		
Restricted log likelihood (constant term only)		
Log likelihood at convergence		
Chi-squared (and associated p-value)	(<	

Estimated Parameter	t Statistic
–0.826	–3.57
0.0000812	6.90
0.000550	7.38
– 0.0600	– 2.73
0.0748	4.54

Poisson-Lognormal Model

- With the negative binomial model, the relationship between the observed crash count and explanatory variables is given as,

$$\lambda_i = EXP\left(\beta x_i + \varepsilon_i\right) = \exp(\boldsymbol{\beta X_i})\exp(\varepsilon_i)$$

where $EXP(\varepsilon_i)$ is a Poisson-lognormal-distributed error term with mean 0 and variance σ_ε^2.

$$\exp(\varepsilon_i)|\sigma_\varepsilon^2 \sim Lognormal(0, \sigma_\varepsilon^2) \text{ 或 } \varepsilon_i|\sigma_\varepsilon^2 \sim Normal(0, \sigma_\varepsilon^2)$$

$$y_i|\lambda_i, \sigma_\varepsilon^2 \sim PLN(\mu_i, \sigma_\varepsilon^2)$$

$$E(Y_i) = \lambda_i \exp(0.5\,\sigma_\varepsilon^2)$$

$$Var(Y_i) = E(Y_i) + [E(Y_i)]^2(\exp(\sigma_\varepsilon^2)-1)$$

Multivariate Poisson-Lognormal Model

- 信号交叉口在不同交通状态下，车辆运行情况不同，由此造成交通冲突发生的概率不同，从而引起单位时间内交通冲突数的差异，所以交通冲突建模应该考虑不同交通状态。
- 相邻两个交通状态之间，单位时间内交通冲突数又并非完全独立，而是相互关联的。多元泊松-对数正态分布回归模型（multivariate Poisson-lognormal model, MVPLN）可以解决不同交通状态下交通冲突存在差异但相互关联的问题。
- 若y_{ik}表示第i个时间段内交通冲突次数属于第k类交通状态，假设y_i服从独立分布，y_{ik}服从参数是μ_{ik}的泊松分布

$$\Pr(Y_{ik}=y_{ik}|\theta_{ik})=\theta_{ik}^{\,y_ik}e^{-\theta_{ik}}/y_{ik}! \qquad i=1,2,\ldots n \qquad k=1,2,\ldots K$$

为处理交通冲突数据离散性，引入误差项ε_{ik}，记：

$$\ln(\theta_{ik})=\ln(\mu_{ik})+\varepsilon_{ik}$$

$$\ln(\mu_{ik})=\beta_{k0}+\beta_{k1}x_{i1}+\ldots\beta_{KJ}x_{iJ}$$

Multivariate Poisson-Lognormal Model

- 信号交叉口在不同交通状态下，车辆运行情况不同，由此造成交通冲突发生的概率不同，从而引起单位时间内交通冲突数的差异，所以交通冲突建模应该考虑不同交通状态。
- 相邻两个交通状态之间，单位时间内交通冲突数又并非完全独立，而是相互关联的。多元泊松-对数正态分布回归模型（multivariate Poisson-lognormal model, MVPLN）可以解决不同交通状态下交通冲突存在差异但相互关联的问题。
- 若y_{ik}表示第i个时间段内交通冲突次数属于第k类交通状态，假设y_i服从独立分布，y_{ik}服从参数是μ_{ik}的泊松分布

$$\Pr(Y_{ik} = y_{ik} | \theta_{ik}) = \theta_{ik}^{y_{ik}} e^{-\theta_{ik}} / y_{ik}! \quad i = 1,2,\ldots n \quad k = 1,2,\ldots K$$

为处理交通冲突数据离散性，引入误差项ε_{ik}，记：

$$\ln(\theta_{ik}) = \ln(\mu_{ik}) + \varepsilon_{ik}$$

$$\ln(\mu_{ik}) = \beta_{k0} + \beta_{k1} x_{i1} + \ldots \beta_{KJ} x_{iJ}$$

Multivariate Poisson-Lognormal Model

式中，x_{ij}表示影响交通冲突的协变量，即解释变量；$\boldsymbol{\beta}_{kj}$是相应解释变量的系数向量；

误差项ε_{ik}服从多维正态分布，即 $\varepsilon_i \sim N_K(0,\Sigma)$。

$$\varepsilon_i=\begin{bmatrix}\varepsilon_{i1}\\ \varepsilon_{i2}\\ \vdots\\ \varepsilon_{iK}\end{bmatrix}\quad \Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1K}\\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2K}\\ \vdots & \vdots & \ddots & \vdots\\ \Sigma_{K1} & \Sigma_{K2} & \cdots & \Sigma_{KK}\end{bmatrix} \tag{5-33}$$

令$\boldsymbol{X}$分别表示协变量矩阵；$\boldsymbol{\beta}$表示相应的系数向量，即$\boldsymbol{\beta}=\{\beta_1, \beta_2,\ldots\beta_J\}$。给定矩阵$(\boldsymbol{X}, \boldsymbol{\beta}, \Sigma)$，$\theta_i$服从独立分布，其概率密度函数为

$$f(\theta_i|X,\beta,\Sigma)=\frac{\exp\{-0.5(\lambda_i^*-\mu_i^*)'\Sigma^{-1}(\lambda_i^*-\mu_i^*)\}}{(2\pi)^{K/2}(\prod_{k=1}^{K}\theta_{ik})|\Sigma|^{1/2}} \tag{5-34}$$

这是一个K维对数-正态分布，式中

$$\theta_i=(\theta_{i1}\ \ \theta_{i2}\ldots\ \ \theta_{iK})',\qquad \theta_i^*=(\ln(\theta_{i1})\ \ \ln(\theta_{i2})\ldots\ \ \ln(\theta_{iK}))'$$

$$\mu_i=(\mu_{i1}\ \ \mu_{i2}\ldots\ \ \mu_{iK})',\qquad \mu_i^*=(\ln(\mu_{i1})\ \ \ln(\mu_{i2})\ldots\ \ \ln(\mu_{iK}))'$$

Zero-Inflated Poisson and Negative Binomial Regression Models

- There are certain phenomena where an observation of zero events during a given time period can arise from two qualitatively different conditions. One condition may result from simply failing to observe an event during the observation period. Another qualitatively different condition may result from an inability to ever experience an event.
- For example, for straight sections of roadway with wide lanes, low traffic volumes, and no roadside objects, the likelihood of a vehicle accident occurring may be extremely small, but still present because an extreme human error could cause an accident.

Zero-Inflated Poisson and Negative Binomial Regression Models

- To address phenomena with zero-inflated counting processes, the zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models have been developed. The ZIP model assumes that the events,
- $Y = (y_1, y_2, \ldots, y_n)$, are independent and the model is

$$y_i = 0 \quad \text{with probability } p_i + (1-p_i)EXP(-\lambda_i)$$

$$y_i = y \quad \text{with probability } \frac{(1-p_i)EXP(-\lambda_i)\lambda_i^y}{y!}$$

- where p_i is the probability of being in the zero state and y is the number of events per period. Maximum likelihood estimates are used to estimate the parameters of a ZIP regression model and confidence intervals are constructed by likelihood ratio tests.

Zero-Inflated Poisson and Negative Binomial Regression Models

- The ZINB regression model follows a similar formulation with events, $Y = (y_1, y_2, \ldots, y_n)$, being independent and

$$y_i = 0 \quad \text{with probability } p_i + (1-p_i)\left[\frac{1/\alpha}{(1/\alpha)+\lambda_i}\right]^{1/\alpha}$$

$$y_i = y \quad \text{with probability } (1-p_i)\left[\frac{\Gamma((1/\alpha)+y)\, u_i^{1/\alpha}(1-u_i)^y}{\Gamma(1/\alpha)y!}\right], \quad y = 1, 2, 3, \ldots$$

- where $u_i = (1/\alpha)\ [(1/\alpha)+\lambda_i\ ]$. Maximum likelihood methods are again used to estimate the parameters of a ZINB regression model.

Zero-Inflated Poisson and Negative Binomial Regression Models

- To test the appropriateness of using a zero-inflated model rather than a traditional model. Vuong proposed a test statistic for models that is well suited for situations where the distributions (Poisson or negative binomial) are specified. The statistic is calculated as (for each observation i)

$$m_i = LN\left(\frac{f_1\left(y_i \mid \mathbf{X}_i\right)}{f_2\left(y_i \mid \mathbf{X}_i\right)}\right)$$

- where $f_1(y_i|X_i)$ is the probability density function of model 1, and $f_2(y_i|X_i)$ is the probability density function of model 2.

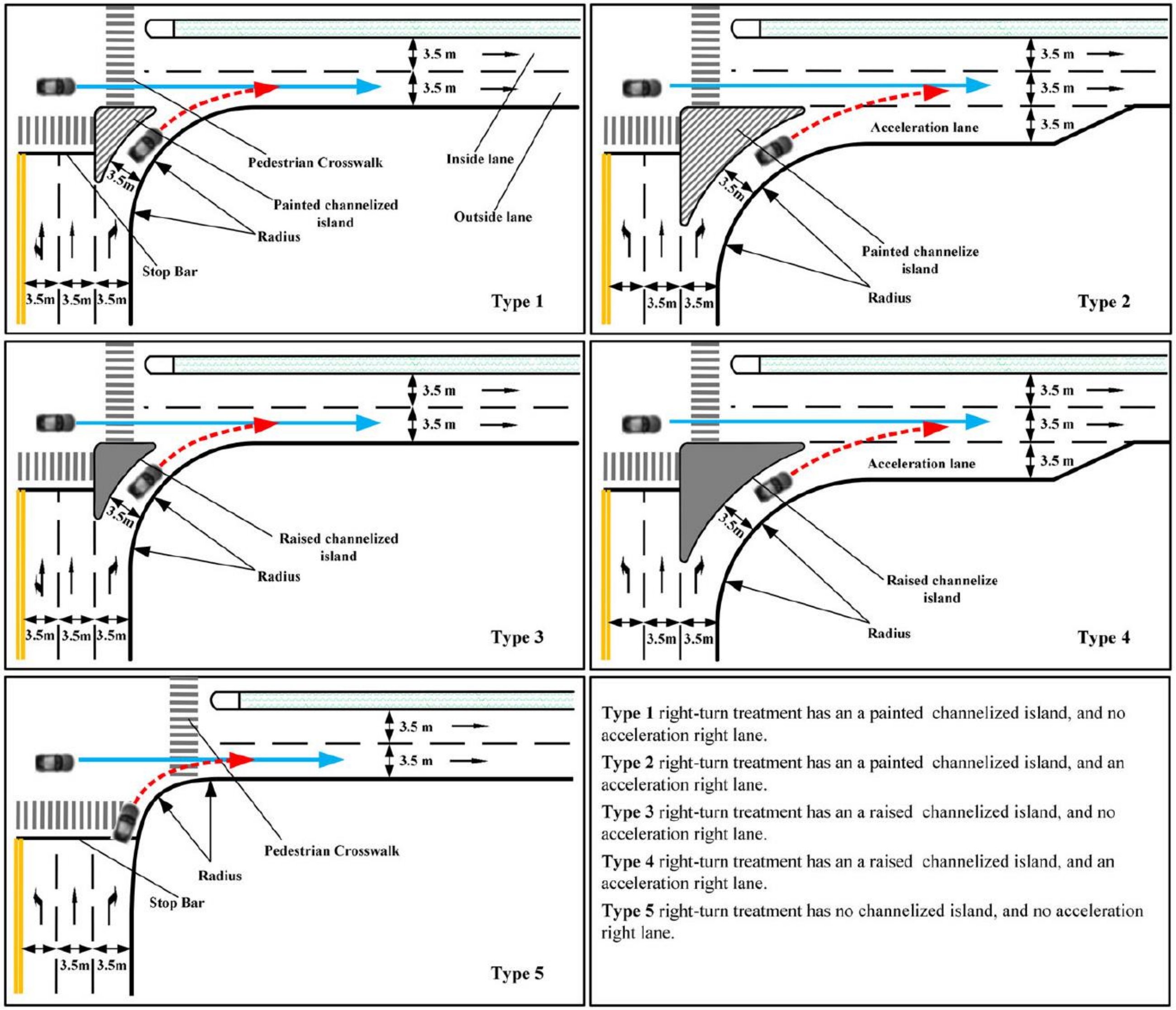
Zero-Inflated Poisson and Negative Binomial Regression Models

- Using this approach, Vuongs' statistic for testing the hypothesis of model 1 versus model 2 is:

$$V=\frac{\sqrt{n}\left[(1/n)\sum_{i=1}^{n}m_i\right]}{\sqrt{(1/n)\sum_{i=1}^{n}(m_i-\overline{m})^2}}=\frac{\sqrt{n}\,(\overline{m})}{S_m}$$

- where $\overline{m}$ is the mean $((1/\text{n})\sum_{i=1}^{n}m_i)$, S_m is standard deviation, and n is a sample size.
- if V is less than $V_{critical}$ (1.96 for a 95% confidence level), the test does not support the selection of one model over another.

Random-parameters Count Models



Random-parameters Count Models

Let Y_i represent the number of conflicts during a specific time period i, and Y_i is assumed to follow a Poisson distribution with parameter λ_i as follows

$$Y_i \sim Poisson(\lambda_i) \tag{1}$$

Where, λ_i is the expected number of conflicts during the specific time period i.

To deal with the over-dispersed issue, it is assumed that

$$\lambda_i = \mu_i \exp(u_i) \tag{2}$$

A logarithm link function that connects μ_i to a linear predictor is given by

$$\mu_i = \exp(\beta_0 + \beta_1 x_1 + \cdots \beta_k x_k) \tag{3}$$

Where, x_1, $x_2, \ldots,$ x_k are explanatory variables; and β_1, $\beta_2, \ldots,$ β_k are model parameters.

The fixed-parameter PLN model is obtained by the following assumption

$$\exp(u_i) \sim Lognormal(0, \sigma_u^2) \tag{4}$$

Where, ${\sigma_u}^2$ is the extra Poisson variance.

Random-parameters Count Models

- The fixed parameter PLN model assumed that the effects of contributing factors to the RTOR conflicts frequency are the same across different right-turn treatments. However, there could be unobserved heterogeneity associated with the RTOR conflicts occurrence across right-turn treatments due to the driver behavior, vehicle type, traffic condition, geometric design, and interaction of them.
- To account for the heterogeneity, the random-parameter model is developed by allowing each of the model parameter follows a specific distribution, implying varying influences of explanatory variables on the expected RTOR conflicts frequency across right-turn treatments. The link function is shown as

Random-parameters Count Models

$$\mu_i = \exp\left(\beta_{i,0} + \beta_{i,1}x_1 + \cdots \beta_{i,k}x_k\right) \tag{5}$$

Where,

$$\beta_{i,j} \sim N\left(\beta_j, \sigma_j^2\right), j = 0, 1, \ldots k, \tag{6}$$

- The random-effect model is developed by adding a location-specific effect to the constant term, implying varying influences of the constant on the expected traffic conflicts frequency at the right-turn treatment. The link function is shown as

$$\mu_i = \exp(\beta_{m,0} + \beta_1 x_1 + \cdots \beta_k x_k) \tag{7}$$

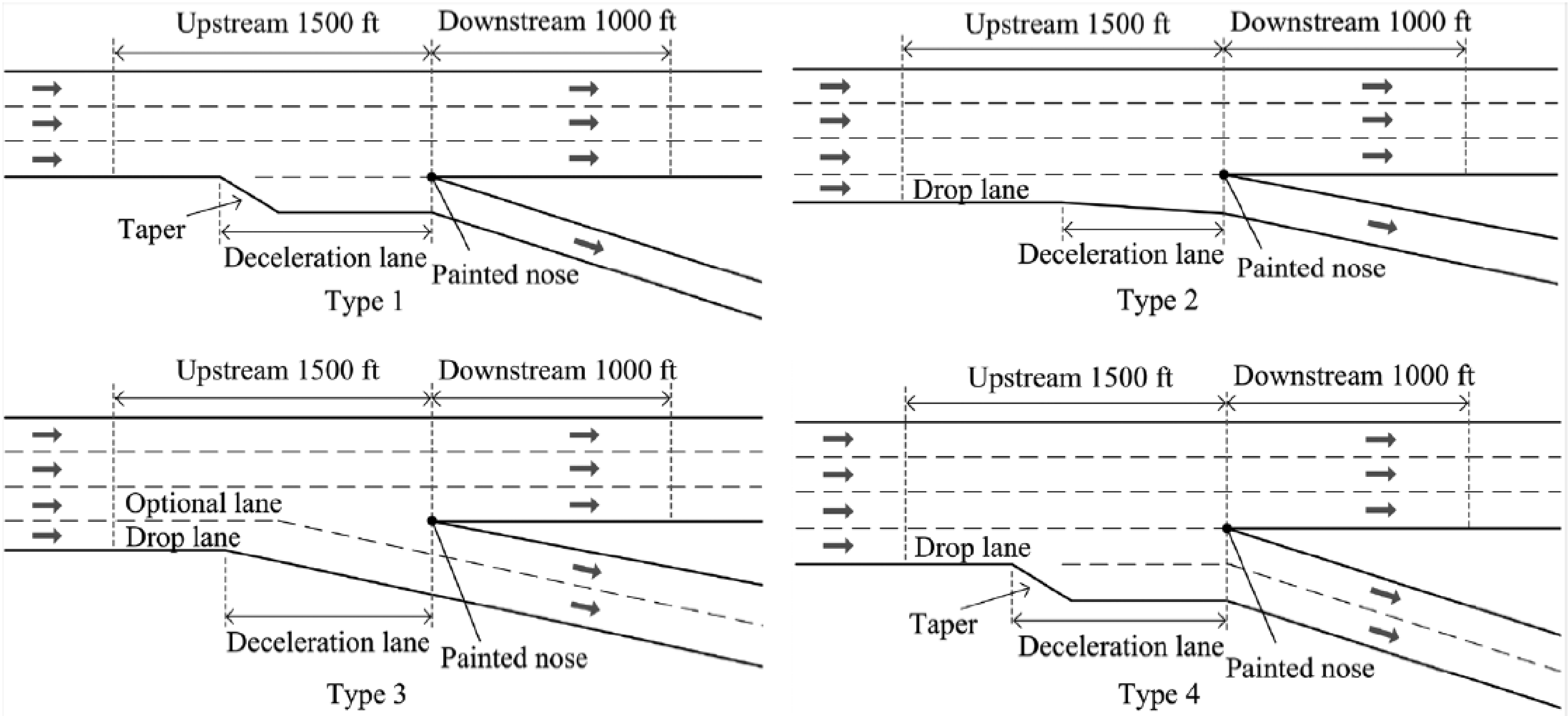
Where,

$$\beta_{m,0} \sim N\left(\beta_0, \sigma_m^2\right) \tag{8}$$

Where, m is the number of right-turn treatments.

Tobit Models

$$CR_j = \frac{10^6 \times N_j}{365 \times T \times L \times \sqrt{AADT_{mainline} \times AADT_{ramp}}} \quad (1)$$



Tobit Models

$$CR_j = \frac{10^6 \times N_j}{365 \times T \times L \times \sqrt{AADT_{mainline} \times AADT_{ramp}}} \quad (1)$$

- Traditional Tobit Model

$$Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + , \cdots , + \beta_k X_{ik} + \varepsilon_i \quad (2)$$

$$Y_i = \begin{cases} Y_i^*, & \text{if } Y_i^* > 0 \\ 0, & \text{if } Y_i^* \leqslant 0 \end{cases}, \ i = 1, 2, \cdots N \quad (3)$$

- where Y_i is the dependent variable (observed values of crash rates), X_{ik} is the kth explanatory variable for observation i, b0 is the model intercept, β_1, β_2, ., β_k are the model parameters, Y_i* is a latent variable observed only when positive， ε_i

$$\varepsilon_i \sim N(0, \sigma^2)$$

Tobit Models

- Random Parameters Tobit Model

$$Y_i^* = \beta_0 + \beta_{i,1}X_{i1} + \beta_{i,2}X_{i2} + , \cdots, + \beta_{i,k}X_{ik} + \varepsilon_i \quad (5)$$

- where the coefficients are set to be random parameters that are assumed to be normallv distributed as following

$$\beta_{i,j} \sim N(\beta_j, \sigma_j^2),\ j = 1, 2, \cdots, k$$

- Grouped Random Parameters Tobit Model
- Assuming that the ith observation belongs to group g(i) ∈{1, 2, 3, 4}, the GRP-Tobit model is given as

$$Y_i^* = \beta_0 + \beta_{g(i),1}X_{i1} + \beta_{g(i),2}X_{i2} + , \cdots, + \beta_{g(i),k}X_{ik} + \varepsilon_i$$

Tobit Models

- Random Effect Tobit Model
- Another way to account for the within-group correlations and the heterogeneity across different types of diverge areas is to allow only the intercept to vary, which leads to the RI-Tobit model. The RE-Tobit model is given as

$$Y_i^* = \beta_{m,0} + \beta_1 X_{i1} + \beta_2 X_{i2} + , \cdots , + \beta_k X_{ik} + \varepsilon_i \quad (9)$$

- where the intercept is set to be a random parameter that follows a normal distribution as follows

$$\beta_{m,0} \sim N(\beta_0, \sigma_0^2), m = 1, 2, 3, 4$$

Concerns in Model Development

- Over-dispersion
- Under-dispersion
- Time-varying explanatory variables
- Temporal and spatial correlation
- Low sample-mean and small sample size
- Injury-severity and crash-type correlation
- Under-reporting
- Omitted-variables bias
- Endogenous variables
- Fixed parameters

Model type	Advantages	Disadvantages
Poisson	Most basic model; easy to estimate	Cannot handle over- and under-dispersion; negatively influenced by the low sample-mean and small sample size bias
Negative binomial/ Poisson-gamma	Easy to estimate can account for over-dispersion	Cannot handle under-dispersion; can be adversely influenced by the low sample-mean and small sample size bias
Poisson-lognormal	More flexible than the Poisson-gamma to handle over-dispersion	Cannot handle under-dispersion; can be adversely influenced by the low sample-mean and small sample size bias (less than the Poisson-gamma), cannot estimate a varying dispersion parameter
Zero-inflated Poisson and negative binomial	Handles datasets that have a large number of zero-crash observations	Can create theoretical inconsistencies; zero-inflated negative binomial can be adversely influenced by the low sample-mean and small sample size bias
Conway–Maxwell–Poisson	Can handle under- and over-dispersion or combination of both using a variable dispersion (scaling) parameter	Could be negatively influenced by the low sample-mean and small sample size bias; no multivariate extensions available to date
Gamma	Can handle under-dispersed data	Dual-state model with one state having a long-term mean equal to zero
Generalized estimating equation	Can handle temporal correlation	May need to determine or evaluate the type of temporal correlation a priori; results sensitive to missing values
Generalized additive	More flexible than the traditional generalized estimating equation models; allows non-linear variable interactions	Relatively complex to implement; may not be easily transferable to other datasets
Random-effects	Handles temporal and spatial correlation	May not be easily transferable to other datasets
Negative multinomial	Can account for over-dispersion and serial correlation; panel count data	Cannot handle under-dispersion; can be adversely influenced by the low sample-mean and small sample size bias
Random-parameters	More flexible than the traditional fixed parameter models in accounting for unobserved heterogeneity	Complex estimation process; may not be easily transferable to other datasets
Bivariate/multivariate	Can model different crash types simultaneously; more flexible functional form than the generalized estimating equation models (can use non-linear functions)	Complex estimation process; requires formulation of correlation matrix
Finite mixture/Markov switching	Can be used for analyzing sources of dispersion in the data	Complex estimation process; may not be easily transferable to other datasets
Duration	By considering the time between crashes (as opposed to crash frequency directly), allows for a very in-depth analysis of data and duration effects	Requires more detailed data than traditional crash-frequency models; time-varying explanatory variables are difficult to handle
Hierarchical/multilevel	Can handle temporal, spatial and other correlations among groups of observations	May not be easily transferable to other datasets; correlation results can be difficult to interpret
Neural network, Bayesian neural network, and support vector machine	Non-parametric approach does not require an assumption about distribution of data; flexible functional form; usually provides better statistical fit than traditional parametric models	Complex estimation process; may not be transferable to other datasets; work as black-boxes; may not have interpretable parameters

[HTML] The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives

D Lord, F Mannering - Transportation research part A: policy and practice, 2010 - Elsevier

Gaining a better understanding of the factors that affect the likelihood of a vehicle crash has been an area of research focus for many decades. However, in the absence of detailed driving data that would help improve the identification of cause and effect relationships with …

ELSEVIER

Contents lists available at ScienceDirect

Transportation Research Part A

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

The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives

Dominique Lord [a,*], Fred Mannering [b]

[a] Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843-3136, United States

[b] School of Civil Engineering, 550 Stadium Mall Drive, Purdue University, West Lafayette, IN 47907-2051, United States

A R T I C L E I N F O

Article history:
Received 10 February 2010
Accepted 26 February 2010

Keywords:
Highway safety
Literature review
Regression models
Count data models

A B S T R A C T

Gaining a better understanding of the factors that affect the likelihood of a vehicle crash has been an area of research focus for many decades. However, in the absence of detailed driving data that would help improve the identification of cause and effect relationships with individual vehicle crashes, most researchers have addressed this problem by framing it in terms of understanding the factors that affect the frequency of crashes – the number of crashes occurring in some geographical space (usually a roadway segment or intersection) over some specified time period. This paper provides a detailed review of the key issues associated with crash-frequency data as well as the strengths and weaknesses of the various methodological approaches that researchers have used to address these problems.