第三章 计数回归

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.
- 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.

Poisson Regression Model

- 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:

$$P(y_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \qquad P(y_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y_i - u_i)^2}{2\sigma^2}\right)$$

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

Poisson Regression Model

- How to remember the probability density function of Poisson distribution

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

$$y_i! \longrightarrow \frac{\lambda_i^{y_i}}{y_i!} \longrightarrow \frac{\lambda_i^{y_i} e}{y_i!} \longrightarrow \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$$

y很惊讶

看到λ背走了他儿子

他"e"了一声

一刀捅死了λ

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 Xi is a vector of explanatory variables and β is a vector of estimable parameters.

Poisson Regression Model

- In this formulation, the expected number of events per period is given by $E[y_i]=\lambda_i=EXP(\beta X_i)$. 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 yi 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	

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

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,

$$\lambda_i = EXP\left(\beta \mathrm{x}_i + \varepsilon_i\right)$$

  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:

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

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

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, $Y = (y_1, y_2, \dots, y_n)$, being independent and

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

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

- 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 nonnested 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 $f_1(\mathrm{y}_i | X_i)$ is the probability density function of model 1, and $f_2(\mathrm{y}_i | X_i)$ is the probability density function of model 2.

Zero-Inflated Poisson and Negative Binomial Regression Models

- Using this approach, Vuongs' statistic for testing the nonnested 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 $((1/\mathrm{n})\sum_{i=1}^{n} m_i)$, 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-Effects Count Models

- In some cases, there may be reason to expect correlation among observations. This correlation could arise from spatial considerations (data from the same geographic region may share unobserved effects), temporal considerations (such as in panel data—where data collected in the same time period could share unobserved effects), or a combination of the two.
- To account for such correlation, random effects and fixed effects models are considered

Random-Effects Count Models

- To consider random effects in a count data model, the Poisson regression model is rewritten as:

$$P(y_{ij}|\mathbf{X}_{ij},\eta_j) = \frac{EXP\left[-EXP(\boldsymbol{\beta}\mathbf{X}_{ij})\,EXP(\eta_j)\right]\left[EXP(\boldsymbol{\beta}\mathbf{X}_{ij})\,EXP(\eta_j)\right]^{y_i}}{y_{ij}\,!}$$

- where λ_{ij} is the expected number of events for observation i belonging to group j (e.g., a spatial or temporal group expected to share unobserved effects), X_{ij} is a vector of explanatory variables, b is a vector of estimable parameters, and η_j is a random effect for observation group j.
- The most common model is derived by assuming η_j are assumed to be randomly distributed across groups such that $EXP(\eta_j)$ is Gamma-distributed with mean one and variance α.

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

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.