Using Poisson and Negative Binomial Regression Models to Measure the Influence of Risk on Crime Incident Counts

By: Eric L. Piza, PhD

Introduction

The majority of RTM studies to-date have tested risk terrain models through logistic regression with shooting incidents as the dependent variable (Caplan et al., 2011; Kennedy et al., 2011). In logistic regression, the dependent variable is dichotomized to represent either the presence ("1") or absence ("0") of a particular feature. In the case of shootings, logistic regression tests the influence of the independent variable(s) (e.g. "risk values") on the presence or absence of *any* shooting incidents. Given the infrequent occurrence of shootings (compared to other crime types), and the fact that most spatial units are unlikely to have more than 1 incident, logistic regression is an appropriate statistical test in such cases. However, for more frequently-occurring crime types, logistic regression may undercount the total number of crimes since multiple incidents are collapsed into a single unit to fulfill the requirements of logistic regression. Such undercounting of incidents may depreciate the validity of the model, particularly by underestimating the predictive capacity of risk terrain models. This brief discusses the use of count regression models, namely Poisson and negative binomial regression, as a method of overcoming the limitations of logistic regression in RTM studies focusing on more frequently-occurring crime types.

Count Regression Models

While the collapsing of data with more than 2 unique values into a dichotomous variable (e.g. "presence" or "absence") allows any dataset to be incorporated in a logistic regression model, such an approach considerably minimizes variance across spatial units. This may lower the statistical power of the model, which may increase the chance of accepting the null hypothesis when a significant relationship exists between the dependent variable and independent variable(s)—a situation commonly referred to as a "Type II" error (Britt & Weisburd, 2010: 313). Said differently, when underpowered, the statistical model may find the dependent and independent variable(s) to be unrelated even when a significant relationship actually exists. Using a technique able to incorporate non-categorical data preserves the statistical power of the analysis, and may be preferable to logistic regression in certain instances.

Many analysts first consider linear regression models when working with non-categorical data. Linear regression, particularly Ordinary Least Squares (OLS) regression, represents one of the most traditional statistical techniques in applied research. However, OLS regression models rest on particular assumptions which oftentimes are not satisfied with criminology data (Maxfield & Babbie, 2001: 404). OLS assumes that the dependent variable is a continuous value, normally distributed (e.g. not skewed), and linearly related to the independent variables (McClendon, 1994). Crime data, in particular, rarely adheres to these assumptions. Most crime incidents are distributed as "rare event counts." Said differently, smaller values are much more common across spatial units than larger values with zero often being the most commonly observed value. Such a distribution violates the aforementioned assumptions of OLS regression.

Poisson and negative binomial regression models are designed to analyze count data. The "rare events" nature of crime counts are controlled for in the formulas of both Poisson and negative binomial regression. However, Poisson and negative binomial regression models differ in regards to their assumptions of the conditional mean and variance of the dependent variable. Poisson models assume that the conditional mean and variance of the distribution are equal. Negative binomial regression models do not assume an equal mean and variance and particularly correct for overdispersion in the data, which is when the variance is greater than the conditional mean (Osgood, 2000; Paternoster & Brame, 1997). Many have noted that criminological data rarely exhibits equal means and variances, leading to the increased popularity of negative binomial regression in contemporary studies of crime (MacDonald & Lattimore, 2010).

Choosing between Poisson and Negative Binomial Regression Models

Choosing between Poisson and negative binomial models depends on the nature of the distribution of the dependent variable. Analysts commonly select negative binomial regression purely because the assumptions of Poisson models are often not observed with social data. However, Poisson distributions are far from nonexistent, with some researchers even observing the presence of both Poisson and negative binomial distributions within the

same study (see, for example, Braga & Bond, 2008). Therefore, analysts should measure the distribution of their data before choosing between Poisson and negative binomial regression. Measuring the distribution of count data is a fairly straightforward process. Particularly, Pearson Chi-Square goodness-of-fit tests can be incorporated along with exploratory Poisson regression models to measure the distribution of the dependent variable. This simple test identifies the distribution of the data and ensures the selection of the correct statistical model.

Table 1 displays the results of a Poisson model with "Risk Value" as the independent variable and the count of Burglary incidents as the dependent variable. The Pearson goodness-of-fit test results (encompassed by a red rectangle) indicate that the distribution of burglary incidents significantly differs for a Poisson distribution, according to the *p* value of 0.000 ("Prob > chi2"), which falls below the standard threshold of 0.05 . Therefore, negative binomial regression is more appropriate for this particular data set.

Table 1: Poisson goodness-of-fit test

```
Poisson regression                              Number of obs   =      14183
                                                LR chi2(1)      =      89.42
                                                Prob > chi2     =     0.0000
Log likelihood = -5664.3864                     Pseudo R2       =     0.0078

------------------------------------------------------------------------------
   Burg_Count |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
    RiskValue |   .2222416   .0228976     9.71   0.000     .1773632     .26712
        _cons |   -2.35222   .0370918   -63.42   0.000    -2.424919   -2.279522
------------------------------------------------------------------------------

. poisgof

        Deviance goodness-of-fit =  8456.543
        Prob > chi2(14181)       =    1.0000

        Pearson goodness-of-fit  =  21558.41
        Prob > chi2(14181)       =    0.0000
```

Interpreting Model Results

The interpretation of results is the same across count regression models types. Model parameters communicate the same information in both Poisson and negative binomial regression models. Therefore, the subsequent discussion relates to the findings of Poisson models as well as negative binomial regression models.

Table 2 displays the results of a negative binomial regression model with "Risk Value" as the independent variable and the count of Burglary incidents as the dependent variable. The effect of the independent variable on the dependent variable can be determined by the regression coefficient, contained under the "Coef." column in Table 2. Since count regression techniques model the *log* of incident counts, the coefficients can be interpreted as follows: for a one unit change in the independent variable, the log of dependent variable is expected to change by the value of the regression coefficient. In the current example, for every one unit increase in a unit's Risk Value, the *log* count of burglaries is expected to increase by approximately 0.239. The statistical significance of the coefficient is displayed by the *p* value, listed under the column "P>|z|." In this example, the *p* value is 0.000, below the standard threshold of 0.05, meaning that the finding is statistically significant.

Table 2: Negative binomial regression results

Negative binomial regression					Number of obs =	14183
					LR chi2(1) =	62.41
Dispersion = mean					Prob > chi2 =	0.0000
Log likelihood = -5312.6065					Pseudo R2 =	0.0058

Burg_Count	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
RiskValue	.2391447	.0302041	7.92	0.000	.1799457	.2983436
_cons	-2.371178	.0456239	-51.97	0.000	-2.460599	-2.281757
/lnalpha	1.397435	.0713965			1.2575	1.53737
alpha	4.044812	.2887856			3.51662	4.652337

Likelihood-ratio test of alpha=0: chibar2(01) = 703.56 Prob>=chibar2 = 0.000

Rather than reporting Poisson or negative binomial results as a regression coefficient, analysts have the option of measuring the effect of the independent variable on the dependent variable through the Incidence Rate Ratio (IRR). The IRR represents the change in the dependent variable in terms of a percentage increase or decrease, with the precise percentage determined by the amount the IRR is either above or below 1. For certain audiences, this may more clearly communicate independent variable influence than the regression coefficients. In Table 3, the IRR for Risk Value (1.27) suggests that burglary counts increased by approximately 27% with every one unit increase in risk. Conversely, an IRR reporting a 27% *decrease* would be written as 0.73 (a value 0.27 less than 1).

Table 3: Negative binomial regression results with reported incidence rate ratios

Negative binomial regression					Number of obs =	14183
					LR chi2(1) =	62.41
Dispersion = mean					Prob > chi2 =	0.0000
Log likelihood = -5312.6065					Pseudo R2 =	0.0058

Burg_Count	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
RiskValue	1.270162	.0383641	7.92	0.000	1.197152	1.347625
_cons	.0933707	.0042599	-51.97	0.000	.0853838	.1021047
/lnalpha	1.397435	.0713965			1.2575	1.53737
alpha	4.044812	.2887856			3.51662	4.652337

Likelihood-ratio test of alpha=0: chibar2(01) = 703.56 Prob>=chibar2 = 0.000

Conclusion

Poisson and negative binomial regression models afford analysts the opportunity to move beyond categorical data in Risk Terrain Modeling projects. These approaches account for the unique distribution of count data and preserve the validity and power of the statistical analysis. Count regression models also afford analysts the opportunity to precisely measure the data distribution through Pearson goodness-of-fit tests to ensure the selection of the correct model type. In addition, Incidence Rate Ratios can be reported to represent the impact of independent variables in terms of a percentage change in the observed crime counts. While this brief demonstrated these techniques in the Stata 12.1 statistical software package, many readily available statistics programs offer similar functionality.

References

Braga, A., and Bond, B**.** (2008). Policing crime and disorder hot spots: A randomized controlled trial. *Criminology, 46 (3)*: 577-607

Britt, C. and Weisburd, D. (2010). Statistical power. In Piquero A. and Weisburd, D. (eds.) *Handbook of Quantitative Criminology.* Springer: New York, NY.

Caplan, J., Kennedy, L., and Miller, J. (2011). Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting. *Justice Quarterly, 25(2)*: 360-381.

Kennedy, L., Caplan, J., and Piza, E. (2011). Risk clusters, hotspots, and spatial intelligence: Risk terrain modeling as an algorithm for police resource allocation strategies. *Journal of Quantitative Criminology, 27(3)*: 339-362.

MacDonald, J. and Lattimore, P. (2010). Count models in criminology. In Piquero A. and Weisburd, D. (eds.) *Handbook of Quantitative Criminology.* Springer.

Maxfield, M. and Babbie, E. (2001). *Research methods for criminal justice and criminology. Third edition.* Wadsworth/Thompson Learning: Belmont, CA.

McClendon, M. (1994). *Multiple regression and causal analysis.* Itasca, IL: F. E. Peacock Publishers.

Osgood, D.W. (2000). Poisson-based Regression Analysis of Aggregate Crime Rates. *Journal of Quantitative Criminology, 16,* 21-44.

Paternoster, R., and Brame, R. (1997). Multiple routes to delinquency? A test of developmental and general theories of crime. *Criminology, 35,* 45-84.