Deal with Excess Zeros in the Discrete Dependent Variable, the Number of Homicide in Chicago Census Tract

Gideon D. Bahn[1], Raymond Massenburg[2]
[1]Loyola University Chicago, 1 E. Pearson Rm 460 Chicago, IL 60611
[2]University of Illinois at Chicago, 1603 W. Taylor St. Rm 717 Chicago, IL 60612

Abstract

When the outcome variable has many missing data, imputation has been a common method. But when the response is the number of homicide and contains many zeros, we have to deal with zeros in different ways. In order to deal with the zero response many methods has been developed. This paper investigates the procedure of choosing the best fitting model when the response variable has many zeros in the case of homicide. Few models are in consideration: generalized linear regression (GLM) with Poisson and with negative binomial distribution, zero-inflated with Poisson (ZIP) and with negative binomial distribution (ZINB), zero-altered with passion (ZAP or hurdle) and negative binomial distribution (ZANB), and finally two-part model. An example from public health research is used for illustration.

Keywords: GLM with Poisson, ZIP model, ZAP (hurdle) model, Negative Binomial, Two-part model, Homicide

1. Introduction

When a researcher has to analyze a real life data, many difficult situations appear unexpectedly. One of the problems is often found in a response variable, Y. This study deals with the problem of many zero responses in the data. Zero response can be treated in many ways. We can consider it as missing data and impute them in the analysis, but in many cases, zero has important meaning in it; no defect in the product line (Lambert, 1992), no accident in a traffic (Miaou, 1994) and no homicide in Chicago census tract. When the zero has its own important meaning in a discrete response, it should be included as is in the data analysis.

In dealing with many zero responses, few statistical methods have been developed. Intuitively, generalized linear modeling (GLM) with Poisson distribution can be considered first when the response is discrete because this model includes zero responses. But often GLM with Poisson distribution has problem with overdispersion, especially due to excessive number of zeros (Hinde and Demetrio, 1998). In order to fit the model, transformation of the response variable can be adapted but not guaranteed to fix the overdispersion problem. Hurdle model (Mullahy, 1986) and Zero-inflated model (Lambert, 1992) with Poisson distribution dealt with excessive zero responses systematically. Later, Two-part model (Duan, et. al., 1983; Olson & Schafer, 2001) has been developed.

In this paper, we are going to review selection procedure in order to find the best fitting model out of many possible models mentioned above. In order to illustrate, Chicago census tract data from 1990 to 1995 is used; the number of homicide relation to foreclosure rate, subprime lending rate and vacancy rate, which has 196 zero responses out of 840 with 11 missing.

2. Model Comparison

Generalized Linear Model

Generalized linear model (GLM) is extended from linear regression model. Basically, there are two important components in GLM; link function and a set of distribution from the exponential family (Agresti, 2007). Among the distributions of the exponential family, the Poisson distribution has a property includes the discrete response variable with zero and positive numbers. The probabilities of Poisson distribution:

$$P(Y=y)=(\mu^{y}*e^{-\mu})/y!,\ y=0,1,2,3\ldots \quad \text{-- 2.1}$$

Where P(Y=y) is the probability when the success, y, occurs
μ is mean, given μ>0 and equal to its variance

The link function can be created according to the researcher's interest. The more the model gets complicated, the link function should reflect the complexity of the model, but often log-link function or logit link function is used due to its simplicity and interpretability of the parameter(s). For this study, we used log function:

$$g(\mu)=\ln \mu=Bo+B1X1+ \text{----} +BpXp \quad \text{-- 2.2}$$

We can write the likelihood function for this model:

$$L(B/y1, y2,\ldots.., yn)=\Pi i\ (\mu^{yi}*e^{-\mu})/yi!,\ i=1\ldots n \quad \text{-- 2.3}$$

Using this function, the model was built with our data;
The dependent variable is the number of homicide in Chicago census tract during 1990 through 1995, and independent variables are foreclosure rate, subprime lending rate and vacancy rate. Since the number of homicide should be directly proportional to the size of the population, it is reasonable to consider log of population as an offset.

There are few procedures we need to take for model-fit, whether the model fits to the data or not. Traditionally, residuals show how the model deviates from the observed data, thus used in order to check model fit. Calculating residuals, we can either use deviance residuals or Pearson residuals. For this present study, we used two ways to check the model fit; the value of deviance residuals divided by degrees of freedom (df) and assumptions of the model, normality and equal variance (homogeneity). If one assumption is not met, another does not need to be checked. For the equal variance, we looked at the graph of deviance (Pearson) residuals versus fitted value. The deviance residuals are calculated two times of loglikelihood function (2.3) for the present model subtracted from the full (saturated) model. Then we divide deviance residuals by df (n-p-1) and compare with one. When this value is close to one, we consider that the model fits the data adequately, but greater or lesser means over or under dispersion respectively (Abraham & Ledolter, 2006). The graph of deviance residuals versus fitted values is used commonly to judge model fit. When the graph shows no pattern and no outlier (deviance residuals within ± 4), the model fits adequately; otherwise, inadequately. According to the pattern of the residual plot; however, we may consider transformations or alternative models to fit.

Table 2.1 indicates that GLM with Poisson distribution and log link function does not adequately fit the data but has overdispersion problem (1.699>1). Overdispersion in Poisson distribution means that the variance is larger than the mean, which causes that the standard errors are counted too small so that the parameters are overestimated when they are not. The presence of overdispersion may be caused by incorrect specification, data clustering, or outliers.

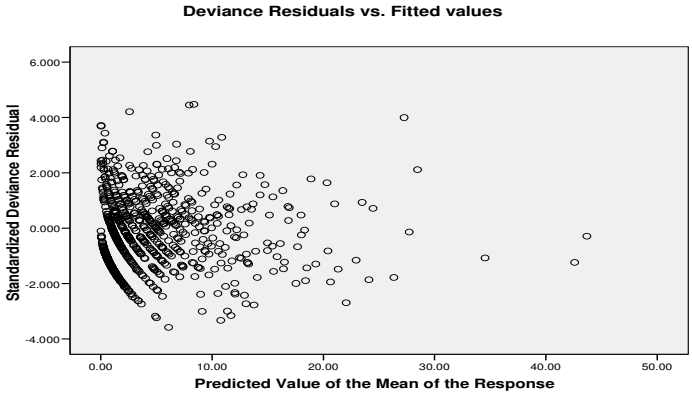
Table 2.1:

	Value	df	Value/df
Deviance	1398.528	823	1.699
Scaled Deviance	1398.528	823	
Pearson Chi-Square	1875.811	823	2.279
Scaled Pearson Chi-Square	1875.811	823	

Log Likelihood(a)	-1740.872		

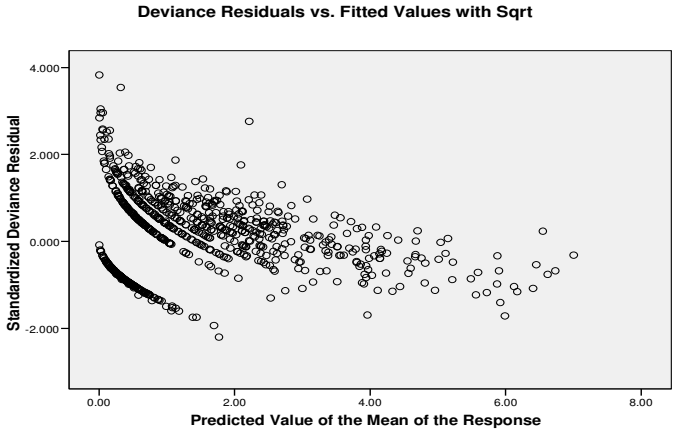
Figure 2.1 shows there are only four outliers have greater than ±4 deviance residuals. But these outliers have small leverage (due to limited space the table is not given) which suggests that they do not influence the model significantly except one outlier. Since residuals are clustering and fanning-in as the predicted values increases, the model violated equal variance assumption. In order to stabilize residuals in GLM we used two ways; transformation of the response variable and changing model from Poisson to negative binomial distribution (Agresti, 2004).

Figure 2.1



First, square-root transformation was done to the response variable, but the model still did not meet the equal variance assumption shown in Figure 2.2. The transformation has no effect on the zero values at all. Even though deviance residuals have become no more than ±4, there is a decreasing pattern with one line separated from other group of dots. Intuitively, we presume that separated line at the bottom of the graph may be caused by zero responses.

Figure 2.2



Secondly, the model with the negative binomial distribution is an alternative way to fix overdispersion problem in Poisson distribution (Greene, 1994, Hinde & Demetrio, 1998).

$$\mathrm{Var}(Yi) = E(Yi) + D\,E(Yi) \qquad \text{-- 2.4}$$

Where D is a dispersion parameter.

When D equals to 0, the GLM with negative binomial distribution is the same as that of the model with Poisson distribution (Agresti, 2007). Therefore, this negative binomial model is nested in the model within

Poisson distribution and the dispersion parameter can be tested for overdispersion of the Poisson model with df=1 (Zorn, 1996).

Figure 2.3

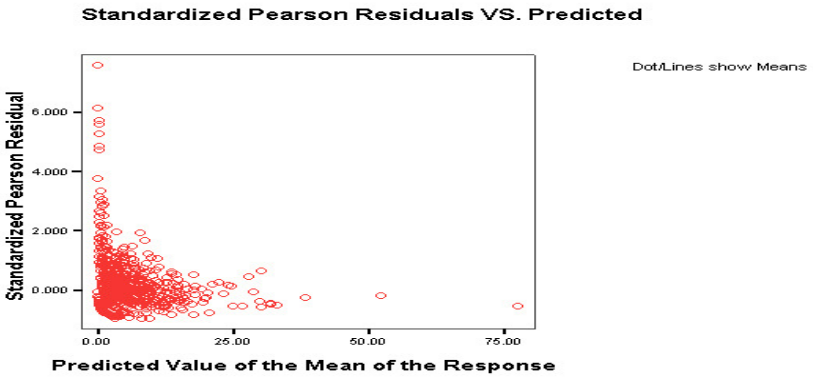


Figure 2.3 suggests that the GLM with negative binomial fixed neither overdispersion problem nor violation of equal variance assumption. It seems that the clustering pattern stems from excessive zero responses. For excessive zero counts, various models have been developed; zero-altered with Poisson (ZAP) by Heilbron (1989), which is introduced as hurdle model by Mullahy (1986), zero inflated Poisson regression (ZIP) by Lambert (1992), and Two-part model by Duan, et. al. (1983) and Olson & Shafer (2001). We are going to investigate each one of these models with our data.

2.2 Zero-altered Poisson (hurdle) Model

Zero-altered Poisson regression is first developed by Mullahy (1986) which was called "hurdle model." The basic idea is that there are two parts of probability; 1) responses that are zeros with probability one, which is similar to ZIP regression model, 2) responses that are from Poisson distribution (asymmetric *Pi=1-e(-e^ Ui*), which is zero-truncated (zero is not included or hurdled over).

$$Pr\,(Yi = y) = \pi o, \quad Yi=0$$
$$Pr\,(Yi = y) = ((1-\pi o)Ui^{yi} * e^{(-Ui)})/yi!(1- e^{(Ui)}), \quad Yi>0 \qquad \text{--2.2.1}$$

2.3 Zero-inflated Poisson Model

ZIP model is originally developed by Lambert (1992) in order to detect the number of defected items in manufacturing equipment. If the equipment is properly aligned, there should be no defect; otherwise, defects may happen. Lambert argued that no defect may come from improperly aligned equipment. In this case, zero is produced systematically. Maybe a certain person, who produced no defect, however, knows how to do it right in improper situation, which makes random zeros. The theory is included this idea in the model as EM algorithm (Lambert, 1992). The number of defects is count data, which is in Poisson distribution, while there are excess zero defects. Therefore, the function of ZIP model follows;

$$Pr(Yi=y) = \pi + (1+\pi)e^{(-u)}, \quad y=0$$
$$Pr(Yi=y) = (1-\pi)e^{(-u)}u^{y}/y!, \quad y>0 \qquad \text{–2.3.1}$$

Lambert's ZIP model has two different points from Mullahy's ZAP model; probability distribution for *Pi* and the distribution of the response estimates, E(*yi*), on *Pi=1*. The probability of distribution for *Pi* assumes symmetric in ZIP model while asymmetric in ZAP model. The distribution of the response estimates on *Pi=1* asserts untruncated Poisson in ZIP model while truncated Poisson in ZAP model. Therefore it is possible to develop a model with the untruncated Poisson distribution of E(yi) on *Pi=1* in the asymmetrical distribution for *Pi* and vice versa.

One of the ways to fit the ZIP model with overdispersion problem is using negative binomial distribution (ZINB) instead of Poisson. We specified the model using the ratio of the variance to the expected value of Y;

$$\mathrm{Var}(yi)/\mathrm{E}(yi)=1+\alpha\mathrm{E}(yi), \text{ Where } \alpha = \ln(б) \quad \text{--2.3.2}$$

In this model when б=1 or α=0, the ratio becomes 1. Therefore, we can test overdispersion problem of the ZIP model through the null hypothesis, Ho: α=0 or б=1. Since ZINB is nested in ZIP, we can use log-likelihood ratio test with df=1. If we reject the null, that means the presence of overdispersion in ZIP model. This means that ZINB model fits better and ZIP.

$$-2(\mathrm{LLzinb}-\mathrm{LLzip}) > 3.841 \text{ with df=1, reject Ho} \quad \text{–2.3.3}$$

2.4 Two-part Model

Two-part model is introduced by Duan et. al. (1983). Out of many possible models, one, two and four part model, they proposed that two-part model fits the best. The method separated samples and has difference analysis. Olson and Schafer (2001) used the method in longitudinal data. In this study, we have divided the samples and analyzed in two ways; logistic regression between zeros and non-zeros and ordinary least square regression. First, we check whether the model fits in logistic regression between zeros and non-zeros in our data. If the model fits in logistic regression within the normality and equal variance assumptions, zero responses are eliminated and use OLS regression without zero responses.

$$\mathrm{Logit}(\pi)=\alpha+B1\mathrm{X1}+B2\mathrm{X2}+B3\mathbf{X}3 \quad \text{--2.3.3}$$

X1=foreclosure, X2=subprime lending **X**3=vacancy rate.

In this section, we proposed five possible models; ZIP, ZINB, ZAP, ZANB and Two-part model. Out of these five models, we have to choose one best-fitting model for our data.

3 Model Comparison and Selection

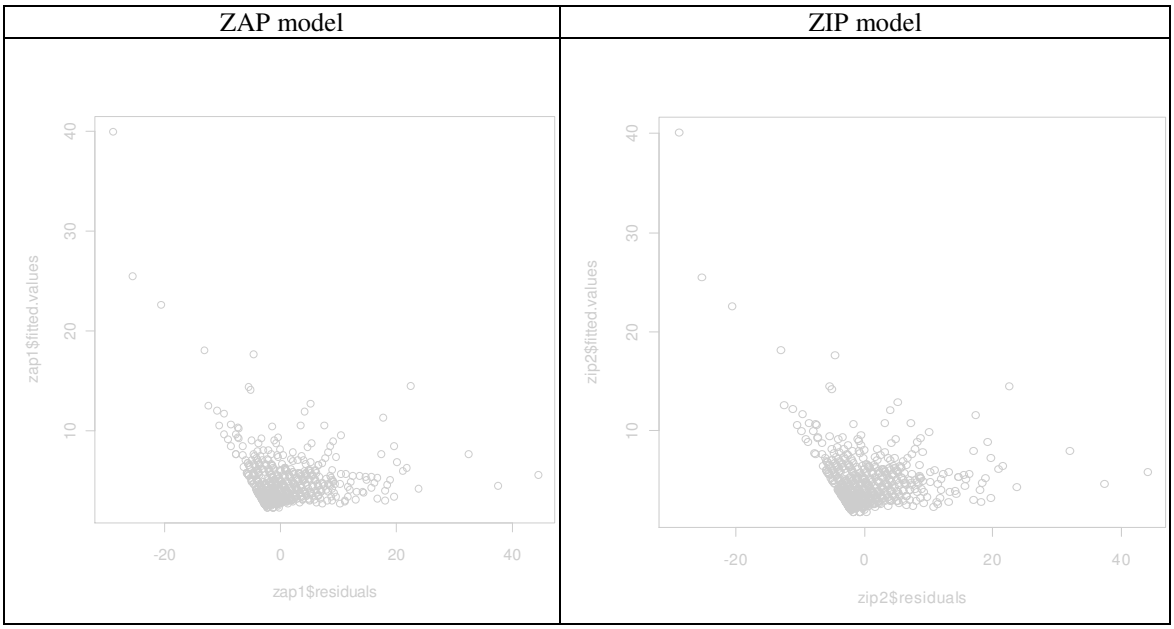
There are many indexes we may use to compare, but an index does not give us clear cut to find which model fits better than the other. Therefore, we decided to use three criteria to compare and select the best-fitting model; 1) comparing log-likelihood ratio, 2) the graph of Pearson residuals versus fitted values and 3) more parsimonious model if two or more models are being compared.

Table 3.1

Model	Log-likelihood ratio
ZAP	2759
ZANB	2094
ZIP	2748
ZINB	2072
Logistic regression	415

By using table 3.1 and figure 3.1, we are able to narrow down to three possible models out of five. Since ZINB and ZANB model are nested into ZIP and ZAP model respectively, we can test overdispersion problem and eliminate two models. Using 2.3.2 calculation, we find that both ZAP and ZIP model have overdispersion problem. In addition, the graphs of the Pearson residuals versus the fitted values, figure 3.1, suggest the same, violating equal variance assumption in both ZAP and ZIP.

Figure 3.1



Out of three models, ZANB, ZINB and logistic regression, usually the model with smaller log-likelihood ratio is considered a better model. But this index is not sufficient enough that logistic regression, which has smaller log-likelihood ratio (LL= 415), is better model than the other models. Therefore, we look at the graphs of the Pearson residuals versus the fitted values in order to find whether each model does not violate equal variance assumption.

Figure 3.2

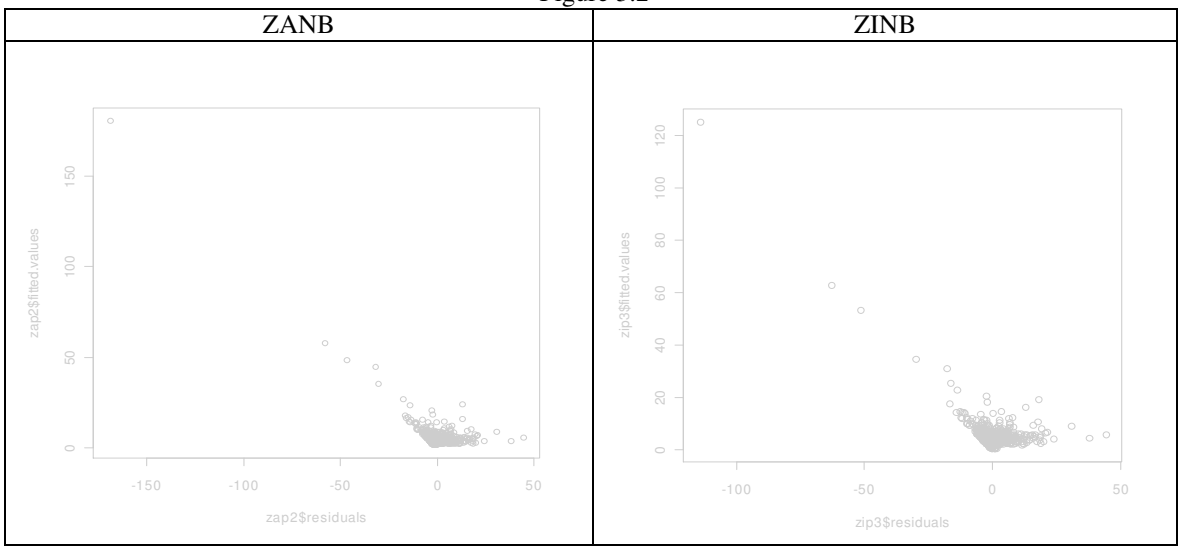


Figure 3.2 asserts that ZANB and ZINB did not fix overdisperson problem from neither ZIP nor ZAP, violating equal variance assumption.

Another way of checking model's fit for logistic regression can be done using Hosmer and Lemeshow test (Abraham & Ledolter, 2006). When the null is retained, the model hold goodness-of-fit. Based on Hosmer and Lemeshow test, the logistic regression model holds goodness-of-fit (Chi-square=5.215, df=8, p=.734). In addition, deviance residual/df=0.968 close to 1 (Agresti, 2007).

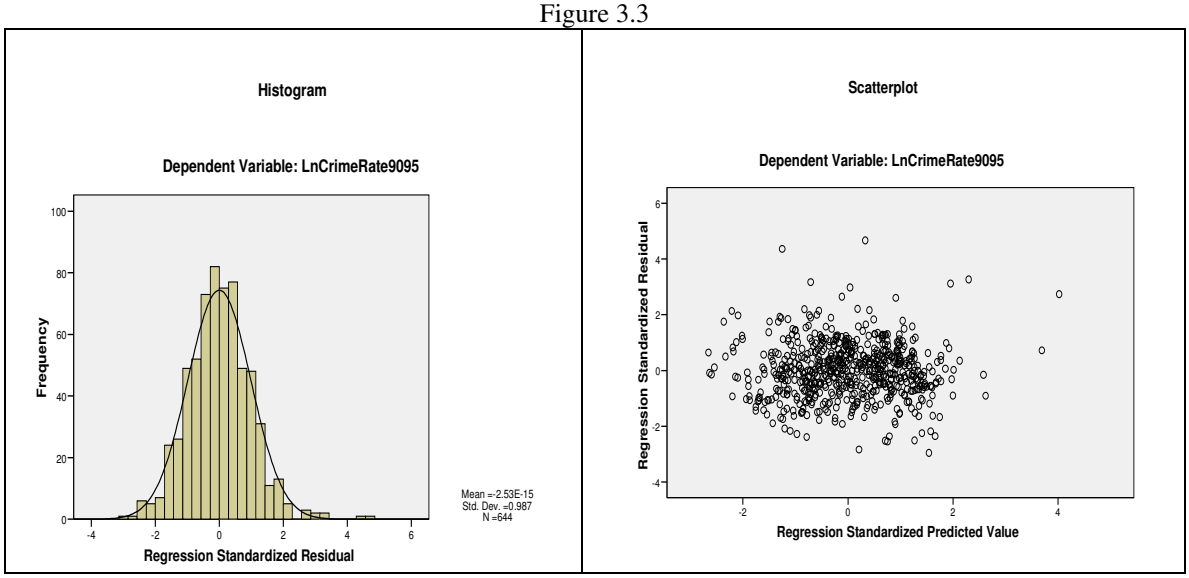
According to the final criterion the logistic model is more parsimonious than two other models, ZANB and ZINB. Yet, we investigate two-part model further without zero responses in OLS regression model. Before we fit the model, we changed the number of homicide into the rate dividing it by the population of each census tract. Then we had log transformation of the rate and the final model as follow;

$$\text{Ln(Homicide)}=\alpha+b1X1+b2X2+b3X3+b4X+ei \quad \text{--3.1}$$

X1=foreclosure, X2=subprime lending **X**3=vacancy rate, X4=demographics, ei=error.

With this model, we have followed few procedures to check the model fit; overall F-test, variance inflation factors (VIF), adjusted R-square, and normality and equal variance assumption. The overall model suggests good fit (F=66.103, p<.0001). All VIFs were less than 10, which suggest no multicollinearity. Adjusted R-square was 0.613, meaning that about 61% of the variation in the response is explained by the OLS regression model with the independent variables. Figure 3.3 shows that the model meets both assumptions, normality and equal variance.

Figure 3.3



The scatter plot shows that there are only two possible outliers, which has bigger than 4 but less than 5 without high leverage values. Eliminating these two points did not make any difference; therefore, they are included in the analysis.

4 Lesson learned and further study

Often the model selection procedure is the agony of most statisticians while analyzing a given real data set. In this study we have learned two lessons; finding possible models with theoretical background and the model selection procedure with precision. Firstly, finding possible models. When the excessive zero responses cannot be imputed, we simply run GLM with Poisson distribution first. After sensing that the model does not fit due to excessive zeros, we start searching for a better fitting model, dealing with the core problem, excessive zeros; instead of trying to fit the data with different transformations. In fact, the transformation does not fix the problem of excessive zeros at all. As a statistician, we know that transformation has nothing to do with zero value, so we honestly should not use the transformation but alternative models in order to fit the data including zero responses. Secondly, in order to select a best-fitting model we should use not only an index to compare models among non-nested models but three criteria; an index (log-likelihood ratio), the residuals vs. fitted values plots and more parsimonious model. These criteria enable us to find the best fitting model for the data set with confidence.

As I mentioned before, there are two more alternative models to explore; ZIP model with asymmetrical probability distribution for *Pi* and ZAP model when probability distribution for *Pi* is symmetrical.

Reference:

Agresti, A. (2007). An Introduction to Categorical Data Analysis (Second ed.), John Wiley & Sons.

Abraham, Bovas & Ledolter, Johannes (2006); Introduction to Regression Modeling; Thomson Brooks/Cole; ISBN 0-534-42075-3.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983), "A Comparison of Alternative Models for the Demand for Medical Care." Journal of Business and Economic Statistics, 1, pp. 115-126.

Greene, W. (1994), Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC-94-10, Department of Economics, New York University.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive Poisson regression model. Journal of Business $ Economic Statistics, 9, 215-222.

Heilbron, D. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data. SIMS Technical Report 9, Department of Epidemiology and Biostatistics, University of Califonia, San Francisco.

Hinde, J. and Demetrio, C. (1998) Overdispersion: models and estimation. Computational Statistics and Data Analysis, 27, 151-170.

Lambert, D. (1992). Zero-inflated Poisson regression, with application to defects in manufacturing. Technometrics 34, 1-14.

Miaou, S.-P. (1994), The relationship between truck accidents and geometric design of road sections. Poisson versus negative binomial regressions. Accident Analysis & Prevention, 26, 471-482.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics, 33, 341-365.

Olsen, M. K. & Shafer, J. L. (2001); A two-part random-effects model for semicontinuous longitudinal data, Journal of the American Statistical Association; June, 96, 454; ABI/INFORM Global pg. 730.

Ridout, M., Demetrio, C. G.B. & Hinde, J. (1998); Models for count data with many zeros; Presented at International Biometric Conference, Cape Town.

Zorn, C. J. W. (1996); Evaluating zero-inflated and hurdle Poisson specifications, Presented at Midwest Political Science Association, Preliminary paper, OH.

Note: To those who attended the presentation, I am sorry that I could not include the interpretability mentioned during the presentation in this paper due to lack of space, which requires few pages to convey the interpretability.