COMPARING THE PERFORMANCE OF DIFFERENT COUNT REGRESSION MODELS

OKAFOR JOSEPH IWEANANDU
NWAGBENU DONMINIC C, KERRY C.C,
KONWE C.S
OJINKEONYE, EBUKA JOACHIN
Department of Mathematics and Statistics, Delta State Polytechnic, Og-washi-Uku

Abstract— In this research, we have considered several regression models to fit the count data encounter in the field of health care provider visit data. We have fitted Poisson (PO), Negative Binomial (NB), Geometric (GEO), Zero-Inflated Poisson (ZIP), Zero-inflated Negative Binomial (ZINB), and Poisson hurdle (PH), Negative Binomial Hurdle (NBH) and Geometric Hurdle (GH) regression models to health care provider visit data. To compare the performance of these models, we analyzed data with moderate percentage of zero counts. Because the variance was less than the mean, we discovered that both GEO and NB models performed better than PO. Also, PH and GH tend to be more superior to PO, ZIP, and ZINB models for the zero inflated and under dispersed count data.

1 INTRODUCTION

BACKGROUND TO THE STUDY

Analyzing data call for determination of the type of data being analyzed. The most basic assumption is that the data follows a normal distribution. However, there are many other types of distributions. The validity of the results can be affected by the dissimilarity between the distribution of the data and the distribution assumed in the analysis. Many outcomes in traffic accident, clinical medicine, and biomedical research are non-negative and discrete in nature B. M. Golam Kibria (2006). Counts are an example of data which does not readily lend itself to the assumption of a normal distribution (Cameron and Trivedi, 1998). Thus it may be natural to model these count data with discrete distribution instead of continuous, which is usually being used as normal. The Poisson (PO) distribution has been used to model the count data for a long time. It has an important constraint that the mean and variance are equal. However, many processes in real life violate the underlying assumption of Poisson (PO) distribution. In that cases the negative binomial (NB) distribution is most preferred and it allows for over-dispersion compared to Poisson distribution. Several researchers have suggested using the NB regression model as an alternative to the PO regression model when the count data are over or under dispersed. Both Poisson and Negative Binomial distribution have been used for predicting the accidents related count frequencies by Miaou (1994 and Lee and Mannering (2002) among others. Unfortunately, the Poisson and NB models do not address the possibility of zero counts and can not fit the data properly. Then corresponding zero augmented models, say zero inflated Poisson (ZIP), zero inflated negative binomial (ZINB), Poisson Hurdle (HP), and Hurdle Negative Binomial (NBH) are very useful to describe the zero inflated and excess zero count data.. The most appropriate reference for ZIP regression model are Hilbe (2011) and Lee et al. (2001) and ZINB regression model is Cameron and Trivedi (1998) among others. The main objective of this thesis is to provide a comprehensive review of these models, discuss how to fit appropriate statistical models for count data using R software and compare these models using Akaike's Information Criterian (AIC), log likelihood and Deviance statistics.

STATEMENT OF THE PROBLEM

Slymen, Ayala, Arredondo, and Elder (2006) found the ZIP and negative binomial ZIP models to be equal. Welsh, Cunningham, Donnelly, and Lindenmayer (2007) found the Hurdle and ZIP models to be equal while Pardoe and Durham (2003) found the negative binomial ZIP model to be superior to both the Poisson and Hurdle models.

One striking characteristic of these articles and others is their differences in terms of the proportion of zeros and the distribution for the non-zeros. Further, the non-zeros varied in terms of their distributions from highly positively skewed to normal to uniform. It is possible that different models yield different results depending on the proportion of zeros and the distribution for the non-zeros.

The best model is the one that appropriately answers the researcher's question. Beyond this, a superior model is one that has close proximity between the observed data and that predicted by the model. In other words, a superior model is one with good fit to the data.

This study compared the fit between the Poisson, ZIP, and Hurdle models as well as their negative binomial formulations. Each analysis will be performed for three different proportions ofzeros and two different amounts of skew for the non-zero distribution. Thus, the intended results would clarify the discrepant findings of previous research.

SIGNIFICANCE OF THE STUDY

The superior model is the appropriate model given the research question. This research provides results that aid researchers in determining the appropriate model to use given zero-inflated data.

RESEARCH QUESTIONS

Model comparisons in this research were based on two measures. One is the deviance statistic, which is a measure of the difference in log-likelihood between two models, permitting a probabilistic decision as to whether one model is adequate or whether an alternative model is superior. This statistic is appropriate when one model is nested within another model. The other measure is Akaike's Information Criterion (AIC). This statistic penalizes for model complexity and permits

comparison of non-nested models; however, it can only be used descriptively. These two measures of model fit were used to compare results from our data and each model was analyzed. Specifically, the measures of model fit were used to answer the following research questions:

i. Given some sets of data, what is the difference in the estimated log-likelihood between (a) the Hurdle model vs. Poisson model?; (b) the Negative binomial Poisson model vs. Poisson model?; c) the Negative binomial Hurdle model vs. negative binomial Poisson model?; (d) the Negative binomial ZIP model vs. ZIP model?; and (e) the Negative binomial Hurdle model vs. Hurdle model?

ii. Given the same set of data, what is the difference in the estimated AIC between all the models?

METHODOLOGY

A good statistical model is the one that provides a good approximate mathematical representation of the data being modeled with particular emphasis being on structure or patterns in the data (Hilbe, 2011). Statistical analysis and modelling of data have become increasingly important in scientific research and study inquiries and the process involves application of appropriate statistical procedure, testing hypotheses, interpreting data results, and coming up with valid conclusions (Clinical Science Research, 2009). In dealing with count data, it make more sense to model these count data using PO, NB, ZIP, ZINB or HURDLE distributions. Regardless of whether the assumed model is a PO, NB, ZIP, ZINB, or HURDLE it will be assumed that the occurrences will be independent of each other.

DATA DESCRIPTION AND ANALYSIS

Data Description

We analyze data on 200 individuals, on their frequency visit to health care provider. The objective is to model the demand for medical care as captured by the number of health care provider visits by the people. To demonstrate the performance of the models, we consider how often people do visit their health provider data which encompass 200 respondents as the dependent variable and the independents variables as If they ever had a problem when they visited an health care provider (x1), their educational level (x2), amount they earn in a month (x3), family size (x4), if they have old people in their household (x5), distance to get to their health providers from their home (x6), if they were currently employed (x7) and type of their primary place of employment; whether is a permanent, temporary or casual job (x8). The summary statistics of the data is shown below:

Table 4.1: Summary Statistics of the Data

	y	x1	x2	x3	x4	x5	x6	x7	x8
% of zero count	6.5	46.5	0	0.5	52	66.5	0.5	29.5	24.5
Mean	2.69	0.54	3.64	2.63	0.55	0.34	2.74	0.71	1
Variance	0.981	0.25	1.992	1.46	0.46	0.224	2.203	0.209	0.905
Skewness	-1.026	-0.141	-0.717	0.572	1.526	0.704	0.251	-0.906	1.841
Std. Error of Skewness	0.172	0.172	0.172	0.172	0.172	0.172	0.172	0.172	0.172
Kurtosis	1.313	-2	-0.598	-0.05	3.455	-1.519	-1.321	-1.191	3.715
Std. Error of Kurtosis	0.342	0.342	0.342	0.342	0.342	0.342	0.342	0.342	0.342

- *OKAFOR JOSEPH IWEANANDU*
  Department of Mathematics and Statistics, Delta State Polytechnic, Ogwashi-Uku
  Email: josephiokafor@yahoo.com
  Phone no: 08032275807
- *NWAGBENU DONMINIC C,*
  Department of Mathematics and Statistics, Delta State Polytechnic, Ogwashi-Uku, Phone no:08036096093
- *KERRY C.C,*
  Department of Mathematics and Statistics, Delta State Polytechnic, Ogwashi-Uku, Phone no:08037743555
- *KONWE C.S,*
  Department of Mathematics and Statistics, Delta State Polytechnic, Ogwashi-Uku, Phone no:08063460810
- *OJINKEONYE, EBUKA JOACHIN, Phone no:08036881079*

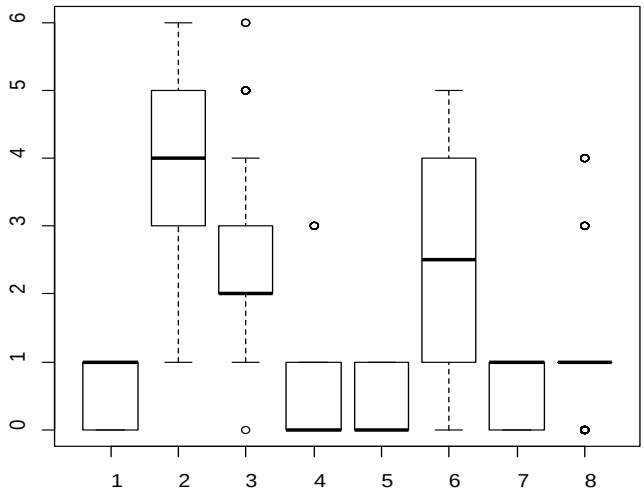


Figure 1: Boxplot of the Explanatory Variables

Data Analysis

Poisson, geometric, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, Poisson hurdle, negative binomial hurdle and Geometric hurdle models were each fit to the data with mixed-effects modeling (MEM), using MASS, car, mhurdle, e.t.c in R 3.0.3 (2014-03-06) on the intent-to-treat sample of all randomized participants. The dependent variable was the count of how often people do visit their health provider. Independent variables were the count of If they ever had any problem(s) when they visited an health provider (x1), their educational level(x2), amount they earn in a month (x3), family size(x4), count of if they have old people in their household (x5), distance to get to their health providers from their home (x6), if they were currently employed (x7) and count of their type of primary place of employment; whether is a permanent, temporary or casual job (x8). The interactions of the reduced variables were included in all the models.

Various statistical tests were applied to evaluate dispersion and compare model fit. Under-dispersion in the Poisson regression was tested by the z statistic. For negative binomial models, the dispersion parameters were tested for difference from zero with z-statistics. To compare goodness of fit between pairs of models, likelihood ratio tests (LR; for full and nested models), Akaike's information criterion (AIC; for non-nested models), and Deviance (for non-nested models) were calculated and used to compare models. After fitting several models, we found that some of these variables are statistically significant to predict the count of how often people do visit their health providers. We have created different models by deleting variables that do not contribute to the model significance for the different count regression method in order to get a best fit model for our data. There are eight sets of count regression data models considered and we have fitted eight different full models for the data set. The R outputs have been provided here.The summaries of statistical analysis have been given in Table 4.3.

4.3 Model Fitting

In Poisson regression model, there are eight explanatory variables (x1, x2, x3, x4, x5, x6, x7, and x8) which do not have significant effect on the health Care provider visits. For geometric and negative binomial regression models, two explanatory variables (x2 and x6) have significant effects on health care provider visits. While for hurdle poisson, hurdle geometric and hurdle negative binomial models, four explanatory variables (x1, x2,x3 and x7) are statistically significant for logit part and none of the variables are significant for the count model part.

The result of Poisson regression analysis is described below

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.43233	-0.32881	0.04342	0.30837	1.25689

Table 4.2: The Parameter Estimates of Selected Poisson Model for the count of health care provider visits

Parameters	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.01383	0.17746	5.713	1.11e-08 ***
x1	0.05456	0.09160	0.596	0.551
x2	-0.05458	0.03764	-1.450	0.147
x3	0.01823	0.04430	0.412	0.681
x4	-0.02704	0.06391	-0.423	0.672
x5	0.05008	0.09987	0.501	0.616
x6	0.04479	0.03264	1.372	0.170
x7	0.01084	0.11041	0.098	0.922
x8	-0.04341	0.05203	-0.834	0.404

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 106.095 on 199 degrees of freedom

Residual deviance: 99.572 on 191 degrees of freedom

AIC: 663.85

Number of Fisher Scoring iterations: 5

We first regressed the response variable 'y' against other explanatory variables viz. 'x1, x2, x3, x4, x5, x6, x7, and x8 in the regression analysis.

The regression equation which is the full model for which the regression equation is now written as:

$$\log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.1$$

On substituting the values of Y and X's, the equation can be written as:

$$\log_e(Y) = 1.01383 + 0.05456X_1 - 0.05458 X_2 + 0.01823 X_3 - 0.02704X_4 + 0.05008 X_5 + 0.04479X_6 + 0.01084X_7 + 0.04341X_8 \quad 4.2$$

Which leads to

$$Y = e^{1.01383} \cdot e^{0.05456X_1} \cdot e^{0.05458 X_2} \cdot e^{0.01823 X_3} \cdot e^{0.02704X_4} \cdot e^{0.05008 X_5} \cdot e^{0.04479X_6} \cdot e^{0.01084X_7} \cdot e^{0.04341X_8} \quad 4.3$$

Table 4.3: Summary of fitted count regression models for our data

Variables	Geometric	Poisson	Negative Binomial	Zero Inflated Poisson	Zero Inflated Neg. Binomial	Negative Binomial Hurdle	Poisson Hurdle	Geometric Hurdle
Intercept	0.997901 (0.111183)	1.01383 (0.17746)	1.001898 (0.110230)	0.994991 (0.170292)	0.991670 (0.172267)	1.12437 (0.19682)	1.124399 (0.196819)	0.81044 (0.37572)
x1	0.056219 (0.056551)	0.05456 (0.09160)	0.055737 (0.056240)	0.052579 (0.088123)	0.049075 (0.092047)	-0.06860 (0.10368)	-0.068602 (0.103681)	-0.09562 (0.19695)
x2	-0.058389 (0.023618)	0.05458 (0.03764)	-0.057480 (0.023408)	-0.051564 (0.038029)	-0.052109 (0.038260)	-0.01553 (0.04228)	-0.015535 (0.042277)	-0.01943 (0.07960)
x3	0.022569 (0.027303)	0.01823 (0.04430)	0.021472 (0.027166)	0.009318 (0.045892)	0.009769 (0.046071)	-0.05422 (0.05173)	-0.054227 (0.051729)	-0.06999 (0.09561)
x4	-0.022376 (0.040126)	-0.02704 (0.06391)	-0.023496 (0.039740)	-0.032761 (0.063379)	-0.032475 (0.063391)	-0.05604 (0.07243)	-0.056048 (0.072427)	-0.07338 (0.13378)
x5	0.062735 (0.062610)	0.05008 (0.09987)	0.059790 (0.062052)	0.059269 (0.096638)	0.056273 (0.099185)	-0.03100 (0.11247)	-0.031000 (0.112470)	-0.04486 (0.21615)
x6	0.050303 (0.020215)	0.04479 (0.03264)	0.048983 (0.020091)	0.045604 (0.031314)	0.044287 (0.032839)	0.01520 (0.03731)	0.015202 (0.037308)	0.02182 (0.07214)
x7	0.006728 (0.068226)	0.01084 (0.11041)	0.007719 (0.067838)	0.014645 (0.110300)	0.13961 (0.12544)	0.139594 (0.12544)	0.17872 (0.23226)	0.006728 (0.068226)
x8	-0.045322 (0.030979)	-0.04341 (0.05203)	-0.044870 (0.031077)	-0.00873 (0.05890)	-0.008736 (0.058902)	-0.01392 (0.10878)	-0.045322 (0.030979)	-0.04341 (0.05203)
Logit part								
Intercept				-30.09 (170.93)	37.526 (377.665)	2.3151 (1.7228)	2.3151 (1.7228)	2.3151 (1.7228)
x1						2.0929 (0.8463)	2.0929 (0.8463)	2.0929 (0.8463)
x2						-0.7878 (0.3234)	-0.7878 (0.3234)	-0.7878 (0.3234)
x3						1.4539 (0.4675)	1.4539 (0.4675)	1.4539 (0.4675)
x4						0.9422 (0.8098)	0.9422 (0.8098)	0.9422 (0.8098)
x5						1.4832 (0.9112)	1.4832 (0.9112)	1.4832 (0.9112)
x6						0.4869 (0.2598)	0.4869 (0.2598)	0.4869 (0.2598)
x7					-4.78 (808.52	-2.7391 (1.3136)	-2.7391 (1.3136)	-2.7391 (1.3136)
x8				6.93 (42.75)	8.784 (94.423)	-0.7321 (0.4527)	-0.7321 (0.4527)	-0.7321 (0.4527)
AIC	877.81	663.85	794.22	666.11	668.09	655.98	653.98	653.98

Deviance	43.615	106.095	59.535					
logL	-429.91	-322.93	-388.11	-323.06	-323.05	-308.99	-308.99	-379.36
F test	2.46	0.79	2.387	-8 4.6378	4.634	22	22.3	19.9
Pr(>F-test)	0.01	0.61	0.018	0.796	0.796	0.1	0.13	0.23
Chisq	1.9146	6.523	2.9567	6.275	6.2837	34.351	34.352	31.66
Pr(>Chisq)	0.9835	0.5889	0.937	0.6173	0.6155	0.04871	0.0487	0.01107
⚭	0.1044129	1	0.1619932					

Table 4.3 Summaries fitted count regression models for our data: coefficient estimates from count models, zero augmented models (both with standard errors in parentheses), maximized log-likelihood, AIC, and deviance statistic. The log-likelihood ratio test of the geometric, PO, NB, ZIP, and ZINB produced virtually identical results which are not significant while hurdle models (PO Hurdle, NB Hurdle and the Geometric Hurdle) also produced same results that say that the models are significant.

4.4 Diagnostics

In diagnostic check, scatter plots are very important in checking the adequacy of the model.

Below are some of the plots

Figure2: Geometric Regression fitted plots

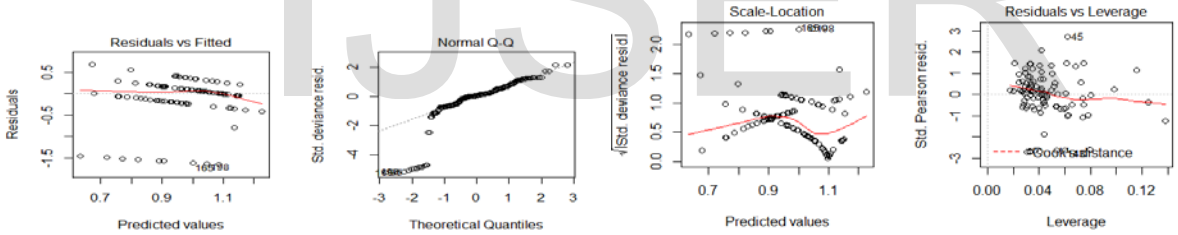


Figure 3: Poisson Regression fitted plots

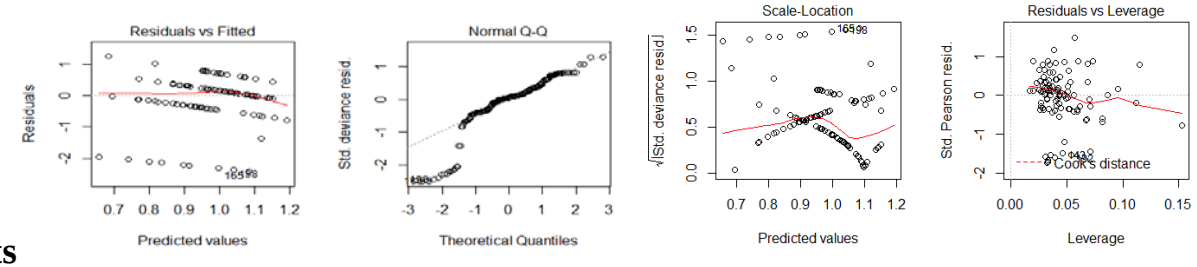
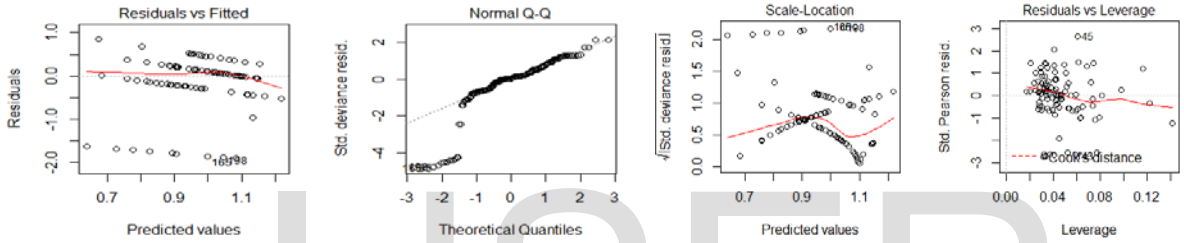


Figure4: Negative Binomial Regression



4.5 Model Comparison

To compare goodness of fit between pairs of models, likelihood ratio tests (for full and nested models), Akaike's information criterion (AIC; for non-nested models), and deviance statistics (for non-nested models) were calculated. AIC recommended Hurdle negative binomial and hurdle geometric regression models as best model, The Deviance statistics suggested the Geometric regression model as the best model while for the log likelihood, the poisson hurdle and the negative binomial hurdle models appeared to best model our data (see table 4.3).

The Poisson model is clearly inferior to all other fits. The geometric and the negative binomial already improves the fit dramatically but can in turn be improved by the hurdle models. This also reflects that the under-dispersion in the data is captured better by the geometric and the negative-binomial-based models than the plain Poisson model. Additionally, it is of interest how the zero counts are captured by the various models. Therefore, the observed zero counts are compared to the expected number of zero counts for the likelihood-based models:

Thus, the Poisson model is again not appropriate whereas the negative-binomial-based models are much better in modeling the zero counts. In summary, the hurdle models lead to the best results (in terms of likelihood) on this data set. For the hurdle model, the zero hurdle components describes the prob-

ability of observing a positive count whereas, for the ZINB model, the zero-inflation component predicts the probability of observing a zero count from the point mass component. Overall, both models lead to the same qualitative results and very similar model fits. Perhaps the hurdle model is slightly preferable because it has the nicer interpretation: there is one process that controls whether a patient sees a health care provider or not, and a second process that determines how many office visits are made.

4.6 Model Selection:

We ran a full glm for each of the our considered count regression model involving count of visit to health provider (y) as our response variables and x1, x2 , x3, ..., x8 as the eight predictor variables. The idea is to find a suitable reduced model if possible that will best fit the model.

For poisson regression

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.43233	-0.32881	0.04342	0.30837	1.25689

Table 4.4: The Parameter Estimates of Selected Poisson Model for the count of health care provider visits

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.01383	0.17746	5.713	1.11e-08
x1	0.05456	0.09160	0.596	0.551
x2	-0.05458	0.03764	-1.450	0.147
x3	0.01823	0.04430	0.412	0.681
x4	-0.02704	0.06391	-0.423	0.672
x5	0.05008	0.09987	0.501	0.616
x6	0.04479	0.03264	1.372	0.170
x7	0.01084	0.11041	0.098	0.922
x8	-0.04341	0.05203	-0.834	0.404

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 106.095 on 199 degrees of freedom

Residual deviance: 99.572 on 191 degrees of freedom

AIC: 663.85

Number of Fisher Scoring iterations: 5

we then performed a backward stepwise regression using *P*-values to delete predictors one-at-a-time. We chose a signific-

ance level α = 0.05 before we started. Then started with the full model, looked at the corresponding model summary, and then identify the predictor which has the largest *P*-value (for the *z* test) aboveour α-level. Then fit a new glm model with that predictor deleted. We used the update() function to achieve this.Furthermore, we looked at the model summary corresponding to the new model, and again identified the predictor for which the *P*-value (for the *z* test) is that predictor deleted, and continue this process until all the remaining P-values were below our α-level. For the poisson regression, throughout the whole process none of the p-values of the predictor variable was less the threshold (0.05).

However, the same process was applied for other count regression models and the reduced models for which the model is significant are shown:

For Geometric regression we have new fitted glm models for the listwise deletion process showm below:

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.3$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 \quad 4.4$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 \quad 4.5$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + \beta_6 X_6 + X_8 \quad 4.6$$

$$log_e(Y) = \beta_0 + \beta_2 X_2 + \beta_6 X_6 + \beta_8 X_8 \quad 4.7$$

$$log_e(Y) = \beta_0 + \beta_2 X_2 + \beta_6 X_6 \quad 4.8$$

And the best reduced model is

$$logY_e = 1.04121 - 0.5599 X_1 + 0.05317 X_2 \quad 4.9$$

For Negative Binomial,

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.10$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 \quad 4.11$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 \quad 4.12$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 \quad 4.13$$

$$log_e(Y) = \beta_0 + \beta_2 X_2 + \beta_6 X_6 + \beta_8 X_8 \quad 4.14$$

$$log_e(Y) = \beta_0 + \beta_2 X_2 + \beta_6 X_6 \quad 4.15$$

And the best reduced model is

$$logY_e = 1.04121 - 0.5599 X_1 + 0.05317 X_2 \quad 4.16$$

Poisson Hurdle Regression

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.17$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.19$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_7 X_7 \quad 4.20$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 \quad 4.21$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad 4.22$$

And the model is

$$logY_e = \begin{cases} \\ 2.1817 + 1.0625X \end{cases} \quad 4.23$$

Negative Binomial Hurdle

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.24$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.25$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.26$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_7 X_7 \quad 4.27$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 \quad 4.28$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad 4.29$$

And the final model is

$$logY_e = \begin{cases} \\ 2.1817 + 1.0625X \end{cases} \quad 4.30$$

Geometric Hurdle

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.31$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.32$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \quad 4.33$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_7 X_7 \quad 4.34$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 \quad 4.35$$

$$log_e(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad 4.36$$

And the final model is

$$logY_e = \begin{cases} \\ 2.1817 + 1.0625X \end{cases} \quad 4.37$$

Another method we employed is in R is the stepwise regression method. We performed the forward stepwise regression, backward stepwise regression, and the combination of both, but R uses the AIC criterion at each step instead of the criteria used before. To use this procedure in the forward direction, we first fit a base model (with one predictor) and a full model (with all the predictors we considered). To fit a base model in our poisson regression, we call it reduced.ypoi

> reduced.ypoi<- We then used ypoi as our full model.

glm(formula=y~1,family=poisson,data=pdataa)

Table 4.5: The Forward stepwise variable Selected Poisson Model for the count of health care provider visits

Variables	Df	Deviance	AIC
+ x8	1	104.03	654.31
<none>		106.09	654.38
+ x2	1	104.30	654.58
+ x6	1	104.70	654.98
+ x7	1	105.29	655.57
+ x1	1	105.39	655.67
+ x5	1	105.55	655.83
+ x4	1	105.56	655.84
+ x3	1	105.75	656.03

Step: AIC=654.31

y ~ x8

	Df	Deviance	AIC
<none>		104.03	654.31
+ x2	1	102.60	654.88
+ x6	1	102.95	655.23
+ x1	1	103.49	655.77
+ x7	1	103.69	655.97
+ x4	1	103.77	656.05
+ x3	1	103.83	656.11
+ x5	1	103.85	656.13

Call: glm(formula = y ~ x8, family = poisson, data = pdataa)

Coefficients:

(Intercept)	x8
1.05366	-0.06798

Degrees of Freedom: 199 Total (i.e. Null); 198 Residual

Null Deviance: 106.1

Residual Deviance: 104 AIC: 654.3

The forward stepwise regression procedure identified the model which included the predictor x8, but not others, as the one which produced the lowest value of AIC. The same process was performed for other models and the results are shown at the appendix.

We used the same procedure in the backward direction; the command is much simpler, since the full model is the base model.

Table 4.6: The Backward Stepwise Variable Selected Poisson Model for the count of health care provider visits

	Df	Deviance	AIC
- x7	1	99.581	661.86
- x3	1	99.741	662.02
- x4	1	99.753	662.03
- x5	1	99.822	662.10
- x1	1	99.927	662.21
- x8	1	100.276	662.56
- x6	1	101.453	663.73
<none>		99.572	663.85
- x2	1	101.668	663.95

Step: AIC=661.86

y ~ x1 + x2 + x3 + x4 + x5 + x6 + x8

	Df	Deviance	AIC
- x3	1	99.747	660.03
- x4	1	99.764	660.04
- x5	1	99.865	660.15
- x1	1	100.010	660.29
- x8	1	100.323	660.60
<none>		99.581	661.86
- x2	1	101.670	661.95
- x6	1	101.752	662.03

Step: AIC=660.03

y ~ x1 + x2 + x4 + x5 + x6 + x8

	Df	Deviance	AIC
- x5	1	99.958	658.24
- x4	1	99.959	658.24
- x1	1	100.152	658.43
- x8	1	100.484	658.76
<none>		99.747	660.03
- x6	1	101.836	660.12
- x2	1	101.842	660.12

Step: AIC=658.24

y ~ x1 + x2 + x4 + x6 + x8

	Df	Deviance	AIC
- x4	1	100.174	656.45
- x1	1	100.319	656.60
- x8	1	100.932	657.21
- x6	1	101.900	658.18
<none>		99.958	658.24
- x2	1	102.259	658.54

Step: AIC=656.45

y ~ x1 + x2 + x6 + x8

	Df	Deviance	AIC
- x1	1	100.56	654.84
- x8	1	101.32	655.60
- x6	1	102.08	656.36
<none>		100.17	656.45
- x2	1	102.50	656.78

Step: AIC=654.84

y ~ x2 + x6 + x8

	Df	Deviance	AIC
- x8	1	101.78	654.07
<none>		100.56	654.84
- x6	1	102.60	654.88
- x2	1	102.95	655.23

Step: AIC=654.07

y ~ x2 + x6

	Df	Deviance	AIC
<none>		101.78	654.07
- x6	1	104.30	654.58
- x2	1	104.70	654.98

Call: glm(formula = y ~ x2 + x6, family = poisson, data = pdataa)

Coefficients:

(Intercept) x2 x6

1.04689 -0.05361 0.04805

Degrees of Freedom: 199 Total (i.e. Null); 197 Residual

Null Deviance: 106.1

Residual Deviance: 101.8 AIC: 654.1

The backward elimination procedure also identified the best model as one which includes only x2 and x6, not others.

Table 4.7: The "both" stepwise variable Selected Poisson Model for the count of health care provider visits

	Df	Deviance	AIC
+ x8	1	104.03	654.31
<none>		106.09	654.38
+ x2	1	104.30	654.58
+ x6	1	104.70	654.98
+ x7	1	105.29	655.57
+ x1	1	105.39	655.67
+ x5	1	105.55	655.83
+ x4	1	105.56	655.84
+ x3	1	105.75	656.03

Step: AIC=654.31

y ~ x8

	Df	Deviance	AIC
<none>		104.03	654.31
- x8	1	106.09	654.38
+ x2	1	102.60	654.88
+ x6	1	102.95	655.23
+ x1	1	103.49	655.77
+ x7	1	103.69	655.97
+ x4	1	103.77	656.05
+ x3	1	103.83	656.11
+ x5	1	103.85	656.13

Call: glm(formula = y ~ x8, family = poisson, data = pdataa)

Coefficients:

(Intercept) x8

1.05366 -0.06798

Degrees of Freedom: 199 Total (i.e. Null); 198 Residual

Null Deviance: 106.1

Residual Deviance: 104 AIC: 654.3

And the selected best models for other GEO, NB, ZIP, ZINB, Hurdle poisson, NB-hurdle and Geometric-Hurdle for each method are shown below:

Geometric Forward Stepwise Selection; the procedure identified the variable x8 and the backward stepwise selection identified x2, x6 and x8 as the best explanatory variable while the combination of both steps also selected x8 as most significant. For the negative binomial, the model reduced to only variable x8 through forward stepwise selection and the backward stepwise selection identified x2, x6 and x8 as the best explanatory variable while the combination of both steps also selected x8 as most significant.The zero inflated negative binomial supported x1 and x2 as best explanatory variable at the negative binomial part and selected x3, x4, x5, x6 and x7 at the logit part for the backward stepwise selection. Negative binomial hurdle supported x3, x2 and x1 for the count model part and also selected x3, x2 and x1 for the zero hurdle model part for the forward stepwise method. The backward stepwise method and the both methods combined for selection also identified x1, x2 and x3 as the best explanatory variables for both the count model part and the zero hurdle part. Even the poisson hurdle and the geometric hurdle also identified x1, x2 and x3 as the best explanatory variables at both the count model part and the zero hurdle model part for all the stepwise methods.

4.7 Interactions

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.65656	-0.09890	0.00457	0.14886	0.66170

Table 4.8: Interaction between Variables for Geometric Model for the Count of Health Care Provider Visits

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.983266	0.644377	3.078	0.00244 **

x1	3.187807	3.276541	0.973	0.33199
x2	-0.805078	0.203240	-3.961	0.00011 ***
x3	0.021304	0.325270	0.065	0.94786
x6	-0.170188	0.276090	-0.616	0.53845
x7	-1.315607	1.658741	-0.793	0.42882
x1:x2	0.002808	0.784897	0.004	0.99715
x1:x3	-1.727720	1.347931	-1.282	0.20169
x2:x3	0.147704	0.082771	1.784	0.07615 .
x1:x6	-2.646111	2.083192	-1.270	0.20576
x2:x6	0.193125	0.080111	2.411	0.01700 *
x3:x6	-0.085281	0.121789	-0.700	0.48475
x1:x7	-2.786038	3.659781	-0.761	0.44757
x2:x7	0.962720	0.465914	2.066	0.04033 *
x3:x7	0.186877	0.693865	0.269	0.78801
x6:x7	0.679713	0.594264	1.144	0.25434
x1:x2:x3	0.192937	0.301343	0.640	0.52288
x1:x2:x6	0.416354	0.512426	0.813	0.41765
x1:x3:x6	1.143734	0.747775	1.530	0.12802
x2:x3:x6	-0.020803	0.030102	-0.691	0.49047
x1:x2:x7	-0.207640	0.901820	-0.230	0.81818
x1:x3:x7	1.640787	1.530817	1.072	0.28533
x2:x3:x7	-0.227074	0.174201	-1.304	0.19418
x1:x6:x7	2.013240	2.158938	0.933	0.35241
x2:x6:x7	-0.362835	0.158883	-2.284	0.02364 *
x3:x6:x7	-0.129635	0.228198	-0.568	0.57074
x1:x2:x3:x6	-0.213501	0.173982	-1.227	0.22149
x1:x2:x3:x7	-0.149056	0.346786	-0.430	0.66788
x1:x2:x6:x7	-0.211678	0.532922	-0.397	0.69172
x1:x3:x6:x7	-0.890007	0.784526	-1.134	0.25822
x2:x3:x6:x7	0.085203	0.056138	1.518	0.13096
x1:x2:x3:x6:x7	0.144470	0.182450	0.792	0.42957

(Dispersion parameter for Negative Binomial (1) family taken to be 0.1074826)

Null deviance: 43.615 on 199 degrees of freedom

Residual deviance: 37.220 on 168 degrees of freedom

AIC: 919.33

Number of Fisher Scoring iterations: 11

Conclusion

Count regression models afford analysts the opportunity to move beyond categorical data in 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.

This thesis provides both methodological and empirical analysis of health care provider visits data. We have fitted several popular count regression models, Poisson (PO), Negative Binomial (NB), Geometric (GEO), Zero-Inflated Poisson (ZIP), Zero-Inflated Negative Binomial (ZINB), Poisson Hurdle (PH), Negative binomial Hurdle (NBH), and Geometric Hurdle (GH) to predict the health care provider visits. We consider moderate (6.5%) to high (65%) number of zeros in the models. A total of twelve different statistical models were fitted in this thesis. All fitted models include significant explanatory variables. Based on deviance and AIC, it appeared that GEOmodel performed better than PO and NB models, followed by NB Model respectively. However, based on the AIC, it also appeared that GH, PH, and NBH model performed better than PO, NB, GEO, ZIP and ZINB models respectively. The empirical study of this thesis revealed that if the under-dispersion and zero-inflation of health care provider visits is found to be moderate to high, GEO, PH and GH models are potential alternatives to PO and ZIP regression models. Poisson regression models serve well under nearly homogeneous condition, while GEO and NB models serve better while data are under dispersed.

REFERENCES

[1] Agresti, A., & Finlay, B. (2002). Statistical methods for the social sciences. (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

[2] Berk, K. N., & Lachenbruch, P. A. (2002). Repeated measures with zeroes. StatisticalMethods in Medical Research, 11

[3] Bonate, P. L. (2001). A brief introduction to Monte Carlo simulation. ClinicalPharmacokinetics, 40

[4] Cameron, A. C., & Trivedi, P. K. (1998). Regression analysis of count data, New York: Cambridge University Press.

[5] Civettini, A. J., & Hines, E. (2005, January). Misspecification effects in zero-inflated negative binomial regression models: Common cases. Paper presented at annual meeting of theSouthern Political Science Association. New Orleans, LA.

[6] Clarke, K. A. (2001). Testing nonnested models of international relations: Reevaluating realism. American Journal of Political Science, 45

[7] Dobbie, M. J., and Welsh, A. H. (2001). Modeling correlated zero-inflated count data. Australian and New Zealand Journal of Statistics, 43.

[8] Famoye, F., & Singh, K. (2006). Zero-inflated generalized Poisson regression model withan application to domestic violence data, Journal of Data Science, 4,

[9] Hilbe J.M (2011). Negative Binomial Regression New York: Cambridge University Press.

[10] Jang, T. Y. (2005). Count data models for trip generation. Journal of TransportationEngineering, 131.

[11] Jung, B. C., Jhun, M., & Lee, J. W. (2005). Bootstrap tests for overdispersion in zero-inflated Poisson regression model. Biometrics, 61.

[12] Kibria B. M. G. (2006) Applications of some discrete regression models for count data, Pak. j. stat. oper. res. Vol.II No.1 2006

[13] Lachenbruch, P. A. (2002). Analysis of data with excess zeroes. Statistical Methods inMedical Research, 11

[14] Miller, J.M. (2007). Comparing Poisson, Hurdle, and Zip Model fit under varying degree of Skew andzero-inflation.

[15] Min, Y., Agresti, A. (2004). Random effects models for repeated measures of zero-inflated count data. (Technical Report 2004-026), Department of Statistics, University ofFlorida, Retrieved May 8, 2006 from second author

[16] Pardoe, I., & Durham, C. A. (2003). Model choice applied to consumer preferences. InProceedings of the 2003 Joint Statistical Meetings, Alexandria, VA, American StatisticalAssociation.

[17] Slymen, D. J., Ayala, G. X., Arredondo, E. M., & Elder, J. P. (2006). A demonstration of modeling count data with an application to physical activity. Epidemiologic Perspectives& Innovations, 3

[18] Warton, D. I. (2005). Many zeros does not mean zero inflation: Comparing the

goodness-of-fit of parametric models to multivariate abundance data. Environmetrics, 16

[19] Cameron A.C., and Trivedi P.K. (1998). Regression Analysis of Count Data, econometric Society Monographs, No. 30, Cambridge University Press.

[20] McCullagh P., and Nelder J.A. (1989). Generalized Linear Models (Second edn). New York: Chapman and Hall. R: A programming Environment for Data Analysis and Graphics. Version 2.12.0 (2010-10-15).J.S. Bridle, "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition," *Neurocomputing – Algorithms, Architectures and Applications,* F. Fogelman-Soulie and J. Herault, eds., NATO ASI Series F68, Berlin: Springer-Verlag, pp. 227-236, 1989. (Book style with paper title and editor)