COMPARING POISSON, HURDLE, AND ZIP MODEL FIT UNDER VARYING DEGREES OF SKEW AND ZERO-INFLATION

By

JEFFREY MONROE MILLER

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2007

To the memory of my grandfather, Rev. Harold E. Cato.

ACKNOWLEDGMENTS

Several people helped make this study possible. I would like to thank my stepfather and mother, Dr. Daniel and Gail Jacobs as well as my father and stepmother, Jerry and Darnelle Miller for their many years of encouragement. I would also like to thank my supervisory committee chair, M. David Miller, for his unyielding guidance, patience, and support. I thank Dr. Jon Morris for the numerous training experiences. Many professors are appreciated for providing the educational foundations for the dissertation topic including Dr. James Algina and Dr. Alan Agresti.

The idea to research zero-inflation was inspired by experiences with data while consulting on projects. To this extent, I thank those clients – Dr. Courtney Zmach and Dr. Lori Burkhead. Undergraduate faculty that I would like to acknowledge for their inspiration and direction include Blaine Peden, Patricia Quinn, and Lee Anna Rasar. Several friends have been a source of encouragement including Matt Grezik and Rachael Wilkerson. Finally, I thank those who made it financially possible to complete this dissertation including my consulting clients, the University of Florida College of Education, the Lastinger Center, and Adsam, LLC.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

COMPARING POISSON, HURDLE, AND ZIP MODEL FIT UNDER VARYING DEGREES OF SKEW AND ZERO-INFLATION

By

Jeffrey Monroe Miller

May 2007

Chair: M. David Miller
Major Department: Educational Psychology

Many datasets are characterized as count data with a preponderance of zeros. Such data are often analyzed by ignoring the zero-inflation and assuming a Poisson distribution. The Hurdle model is more sophisticated in that it considers the zeros to be completely separate from the nonzeros. The zero-inflated Poisson (ZIP) model is similar to the Hurdle model; however, it permits some of the zeros to be analyzed along with the nonzeros. Both models, as well as the Poisson, have negative binomial formulations for use when the Poisson assumption of an equal mean and variance is violated.

The choice between the models should be guided by the researcher's beliefs about the source of the zeros. Beyond this substantive concern, the choice should be based on the model providing the closest fit between the observed and predicted values. Unfortunately, the literature presents anomalous findings in terms of model superiority.

Datasets with zero-inflation may vary in terms of the proportion of zeros. They may also vary in terms of the distribution for the nonzeros. Our study used a Monte Carlo design to sample 1,000 cases from positively skewed, normal, and negatively skewed distributions with proportions of zeros of .10, .25, .50, .75, and .90. The data were analyzed with each model over

2,000 simulations. The deviance statistic and Akaike's Information Criterion (AIC) value were used to compare the fit between models.

The results suggest that the literature is not entirely anomalous; however, the accuracy of the findings depends on the proportion of zeros and the distribution for the nonzeros. Although the Hurdle model tends to be the superior model, there are situations when others, including the negative binomial Poisson model, are superior. The findings suggest that the researcher should consider the proportion of zeros and the distribution for the nonzeros when selecting a model to accommodate zero-inflated data.

CHAPTER 1
INTRODUCTION

Analyzing data necessitates 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. As such, it is imperative that the researcher choose a method for analyzing the data that maintains a distribution similar to that of the observed data.

Counts are an example of data which does not readily lend itself to the assumption of a normal distribution. Counts are bounded by their lowest value, which is usually zero. A regression analysis assuming a normal distribution would permit results below zero. Further, counts are discrete integers while the normal distribution assumes continuous data. Finally, counts often display positive skew such that the frequency for low counts is considerably higher than the frequencies as the count levels increase.

It is not uncommon to find count data analyzed in a more appropriate manner than assuming a normal distribution. Typically, more appropriate analysis includes specification of a Poisson distribution with a log link, rather than a normal distribution with a Gaussian link. However, this does not guarantee accurate and valid results as other features of the data may warrant an even more sophisticated model.

An example of data requiring a more rigorous treatment of the data is the case of zero-inflation. In this scenario, there are far more zeros than would be expected using the Poisson distribution. As such, a number of methods including the zero-inflated Poisson (ZIP) model and the Hurdle model are available. Further, there are negative binomial variations of these for use when particular assumptions appear to be violated. The choice between the models depends on

whether the researcher believes the zeros are all a complete lack of the quantity being measured or that at least some of the zeros are purely random error.

Statement of the Problem

The results from both simulated and actual data sets in the zero-inflation literature are in much disagreement. Lambert (1992) found the ZIP model to be superior to the negative binomial Poisson model, which was superior to the Poisson model. Greene (1994) found the negative binomial Poisson model to be superior to the ZIP model, which was superior to the Poisson model. Slymen, Ayala, Arredondo, and Elder (2006) found the ZIP and negative binomial ZIP models to be equal. Welsh, Cunningham, Donnelly, and Lindenmayer 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 nonzeros. Some research (Böhning, Dietz, Schlattmann, Mendonça, and Kirchner, 1999) analyzed data in which the proportion of zeros was as low as .216 while others (Zorn, 1996) used proportions as high as .958. Further, the nonzeros 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 nonzeros.

Rationale for the Study

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 was performed for five different proportions of

zeros and three different amounts of skew for the nonzero distribution. The intended results would clarify the discrepant findings of previous research.

Purpose and Significance of the Study

The primary purpose of this study was to determine superiority of fit for various models under varying proportions of zero-inflation and varying levels of skew. As such, determination can be made as to which model has better fit given data with a particular proportion of zeros and a particular distribution. The secondary purpose was to elucidate the reasons for discrepant findings in previous research.

The superior model is the appropriate model given the research question. However, there are situations in which the appropriate model is unknown or
unclear. Further, there may be situations in which a simpler model such as the Poisson may be used in lieu of the more sophisticated Hurdle and ZIP models. 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 nonnested models; however, it can only be used descriptively. These two measures of model fit were used to compare results from data simulations where each dataset included 2,000 cases and each model was analyzed 1,000 times. Specifically, the measures of model fit were used to answer the following research questions:

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated log-likelihood between a) the Negative binomial Poisson model vs. Poisson model; b) the Hurdle model vs. Poisson model?; c) the Negative binomial Hurdle model vs. negative binomial Poisson model?; d) the Negative binomial Hurdle model vs. Hurdle model; and, e) the Negative binomial ZIP model vs. ZIP model?

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated AIC between all models?

CHAPTER 2
REVIEW OF THE LITERATURE

Zero-Inflated Count Data

Count Data

As the name implies, count data is data that arises from counting. They are the "realization of a nonnegative integer-valued random variable" (Cameron & Travedi, 1998, p.1). As such, the response values take the form of discrete integers (Zorn, 1996). Although the lower boundary can feasibly be any integer, it is usually the case that its value is zero. Strictly speaking, there can be no nonnegative numbers. Hence, the data are constrained by this lower bound of zero and no upper bound.

Acknowledgment of concerns over zero-inflation, ignoring covariates, likely dates to Cohen (1954). Cameron and Triverdi (1989, p.10-11) identified many areas in which special models have been used to analyze count data including "models of counts of doctor visits and other types of health care utilization; occupational injuries and illnesses; absenteeism in the workplace; recreational or shopping trips; automobile insurance rate making; labor mobility; entry and exits from industry; takeover activity in business; mortgage prepayments and loan defaults; bank failures; patent registration in connection with industrial research and development; and frequency of airline accidents . . . as well as in many disciplines including demographic economics, in crime victimology, in marketing, political science and government, [and] sociology". Surprisingly, there was no mention of research in education. Examples of variables in educational research that yield count data include a student's number of days absent, number of test items scored correct or incorrect, and number of referrals for disciplinary action.

The lower bound constraint of zero presents the biggest obstacle toward analyzing count data when assuming a normal distribution. It is common for count data to have a skewed distribution that is truncated at the lower bound.

$$skew(Y) = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^3}{(N-1)s_Y{}^3} \tag{2-1}$$

Hence, the data are heteroscedastic with variance increasing as the count increases. Therefore, standard models, such as ordinary least squares regression, are not appropriate since they assume that the residuals are distributed normally with a mean of zero and a standard deviation of one (Slymen, Ayala, Arredondo, & Elder, 2006). Cameron and Triverdi (1998) clarify that the use of standard OLS regression "leads to significant deficiencies unless the mean of the counts is high, in which case normal approximation and related regression methods may be satisfactory" (p.2).

An example of a count data variable is the number of household members under the age of 21 reported by respondents in the Adult Education for Work-Related Reasons (AEWR) survey administered by National Council for Educational Statistics in 2003 (Hagedorn, Montaquila, Vaden-Kiernan, Kim & Chapman, 2004). The sample size was 12,725. This variable has a lower count boundary of zero and an upper count boundary of six. The count distribution is positively skewed at 1.971. The distribution mean of 0.54 is certainly not an accurate measure of central tendency; the median and mode are both zero (i.e., the lower-bound itself), and the standard deviation of 0.999 permits negative values in the lower 68% confidence interval.

Zero-Inflation

It is not uncommon for the outcome variable in a count data distribution to be characterized by a preponderance of zeros. As Tooze, Grunwald, & Jones (2002, p.341) explain,

> Typically, [for count data] the outcome variable measures an amount that must be non-negative and may in some cases be zero. The positive values are generally skewed, often extremely so . . . Distributions of data of this type follow a common form: there is a spike of discrete probability mass at zero, followed by a bump or ramp describing positive values.

The occurrence is primarily in the case of interval/ratio count data and sometimes ordinal data (Böhning, Dietz, Schlattmann, Mendonça, & Kirchner, 1999). Regarding continuous data, Hall and Zhang (2004) explain that these distributions "have a null probability of yielding a zero . . . there is little motivation for a model such as [zero-inflated] normal, because all observed zeros are unambiguous . . . " (p.162). If continuous zeros are inflated and those zeros are of concern, they can be analyzed separately from the nonzeros. The null probability of continuous zeros is evident in measures such as height and age.

The condition of excessive zeros is known as zero-inflation (Lachenbruch, 2002) or as a probability mass that clumps at zeros (Tooze, Grunwald, & Jones, 2002). It has been recognized as an area of research in the mid-60's[1] (Lachenbruch, 2002) when Weiler (1964) proposed a method for mixing discrete and continuous distributions. Min and Agresti (2005) formally define zero-inflation as "data for which a generalized linear model has lack of fit due to disproportionately many zeroes" (p.1). There are simply "a greater number of zero counts than would be expected under the Poisson or some of its variations" (Zorn, 1996, p.1).

The Sources of Zero-Inflation

The zeros can be classified as being either true zeros or sampling zeros. True zeros represent responses of zero that are truly null. Suppose an educational inventory item states "How many college preparatory workshops have you attended?" Some of the respondents in the sample may have no intentions to apply for college. Hence, the number of preparatory

[1] Alternatively, if a scale is bound, it is reasonable to consider an inflated upper bound. In this case, scale reversal and subsequent appropriate analysis if justified (Lachenbruch, 2002).

workshops attended may never be greater than zero. Sampling zeros, on the other hand, arise as a probability. There are a proportion of college-bound students who have not attended a workshop due to the possibility that the workshop was not (or is not yet) available. Alternatively, some college-bound students may feel prepared and have no reason to participate in a workshop.

Hence, the mechanism underlying zero-inflation can arise from one or both of 1) a possibility that no other response is probabilistic, or 2) that the response is within a random sample of potential count responses. Martin, Brendan, Wintle, Rhodes, Kuhnert, Field, Low-Choy, Tyre, and Possingham (2005) term the sampling zeros as 'false zeros' and include error as a source of zeros. They state, "Zero inflation is often the result of a large number of 'true zero' observations caused by the real . . . effect of interest . . . However, the term [zero-inflation] can also be applied to data sets with 'false zero' observations because of sampling or observer errors in the course of data collection" (p.1235).

Often, the data contains both types of zeros. This is the result of a dual data generating process (Cameron & Trivedi, 1998). For example, some adults in the AEWR sample may have had true-zero household members under the age of 21 because they are unable to bear children or desire to bear children. Alternatively, they may have random-zero household members under the age of 21 because these adults do have such children but not as members of the household.

Impact of Zero-Inflation on Analyses

"Much of the interest in count data modeling appears to stem from the recognition that the use of continuous distributions to model integer outcomes might have unwelcome consequences including inconsistent parameter estimates" (Mullahy, 1986, p.341). In the typical count data scenario, the zero left-bound implies heteroscedasticity (Zorn, 1996). An even greater problem with zero-inflated distributions, beyond this inadequacy of analyzing such a skewed and heteroscedastic distribution as if it were normal (Tooze, Grunwald, & Jones, 2002) is that they

yield "surprisingly large inefficiencies and nonsensical results" (King, 1989, p126). Martin, Brendan, Wintle, Rhodes, Kuhnert, Field, Low-Choy, Tyre, and Possingham (2005) and McCullagh and Nelder (1989) explain that zero-inflation is a special case of overdispersion in which the variance is greater than it should be given a particular distributional shape and measure of central tendency. The impact is biased/inconsistent parameter estimates, inflated standard errors and invalid inferences (Jang, 2005; Martin, Brendan, Wintle, Rhodes, Kuhnert, Field, Low-Choy, Tyre, and Possingham, 2005).

Simple Solutions to Zero-Inflation

Deleting zeros

The simplest of solutions is to delete all cases having responses of zero on the variable of interest. A large proportion of total responses would then be removed from the total dataset. This would then result in a loss of valuable information impacting statistical conclusion validity (Tooze, Grunwald, & Jones, 2002). The sample size may also then be too small for analyses of the non-zero values.

Assuming normality

Another simple solution is to ignore the zero-inflation, assume asymptotic normality, and analyze the data using standard techniques such as ordinary least squares regression.

$$hhm_i = \beta_0 + \beta_1 Sex_i + \beta_2 Age_i + \varepsilon_i \quad (2\text{-}2)$$

According to this model, the number of household members under the age of 21 for adult respondent *i* is predicted from the overall mean, a coefficient relating the respondent's sex to *hhm*, a coefficient relating the respondent's age to *hhm*, and error. The model assumes that the residuals for *hhm* are distributed normally with a mean of zero and a common variance, σ^2. For the first equation, **y** is a vector of responses, **X** is a design matrix for the explanatory variable

responses, β is a vector of regression coefficients relating **y** to **X**, and ε is a vector of residuals measuring the deviation between the observed values of the design matrix and those predicted from the fitted equation.

Transforming Zeros

Another simple solution is to transform the counts to coerce a more normal distribution (Slymen, Ayala, Arredondo, & Elder, 2006). Since count distributions often appear to be positively skewed, one reasonable transformation involves taking the natural logarithm of the responses to the predictor variables. However, assuming the zeros haven't been deleted, the transformation will not work since the natural logarithm of zero is undefined (Zhou & Tu, 1999; King, 1989).

Sometimes natural log transformations for zero are handled by adding a small value, such as .001, to the zeros. However, this then leads to an inflation of that transformed adjusted value. If 70% of the scores are zero, the resulting transformed distribution will have a 70% abundance of the transformed value (Delucchi & Bostrom, 2004).[2] Further, since the transformation is linear, this technique has been shown to yield biased parameter estimates that differ as a function of the adjustment quantity (King, 1989). Although the undefined log zero problem has been handled, the original problems pervade. As Welsh, Cunningham, Donnelly, & Linenmayer (1996) state, "It is clear for data with many zero values that such an approach will not be valid as the underlying distributional assumptions (linearity, homoscedasticity and Gaussianity) will [still] be violated" (p.298). Finally, for any technique, transformations sometimes create a new problem while solving the old one; "a transform that produces constant variance may not produce normality . . . (Agresti, 1996, p.73)".

[2] This implies then that, beyond the dual generating process for zeros, the problem can be generalized from inflated zeros to inflated lower boundaries for count data.

Generalized Linear Models

Bryk and Raudenbush (1996) state, "There are important cases . . . for which the assumption of linearity and normality are not realistic, and no transformation can make them so" (p.291). As previously demonstrated, count data is likely to be one such case. Instead of deleting cases or transforming the data, it is more reasonable to specify a different distribution. As explained by Hox (2002), although it is nice to be able to transform data, "modeling inherently nonlinear functions directly is sometimes preferable, because it may reflect some 'true' developmental process" (pp. 93-94). In order for a model to be 'inherently nonlinear' (Hox, 2002), there must be no transformation that makes it linear.[3] These nonlinear models belong to the class of generalized linear models (GLM).

The following explanation of generalized linear models based on the seminal work of McCullagh and Nelder (1989) with additional clarification by Lawal (2003) and Agresti (1996). Lawal (2003) explains that generalized linear models are a subset of the traditional linear models that permit other possibilities than modeling the mean as a linear function of the covariates. All GLM possess a random component, a systematic component, and a link function. As explained by Agresti (1996), the random component requires the specification of the distribution for the outcome variable. One could specify this distribution to be normal; hence, classical models such as ordinary least squares regression and analysis of variance models are included within this broader class of generalized linear models. Other possible random components that could be specified include the binomial distribution, negative-binomial distribution, gamma distribution, and Poisson distribution. Specifying the random component depends on the expected population distribution of the outcome variable. Given both zero-inflation and truncated count data yielding

[3] Transforming covariates (e.g., including polynomial terms) may graphically appear to be nonlinear while still be linear in the parameters (Singer & Willett, 2003).

an odd-shaped skewed distribution, the random component plays an important part in obtaining valid results.

In order to better understand the formulation of the three components, it is necessary to clarify the theoretical foundations of distributions. The probability density function for a normal distribution is

$$f(y;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(y-\mu)^2}{2\sigma^2}), \tag{2-3}$$

which, given random variable $X \sim N(\mu,\sigma^2)$, reduces to the standard normal probability density function

$$f(y)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2}{2}), \tag{2-4}$$

which when transformed to the cumulative density function yields

$$\Phi(y)=F(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}\exp(-\frac{u^2}{2})du. \tag{2-5}$$

A convenient method for obtaining the parameters is to use the distribution's moment generating function (Rice, 1995). For the normal distribution, this function is

$$M_y(t)=E[\exp(tY)]=\exp(\mu t+\frac{\sigma^2t^2}{2}). \tag{2-6}$$

The logarithm of the moment generating function yields the cumulant generating function, which then yields the moments of the distribution. For the normal distribution, the first moment is the mean (μ), and the second moment is the variance (σ^2).

Strictly speaking, a requirement for GLM is that the outcome has a distribution within the exponential family of models (EFM) (McCullagh and Nelder, 1996). These distributions are

defined primarily by a vector of natural parameters (θ) and a scale parameter (ϕ). The formulation is given by

$$f_Y(y;\theta,\phi) = \exp\{(\frac{(y\theta - b(\theta))}{a(\phi)}) + c(y,\phi)\}. \quad (2\text{-}7)$$

At first glance, it seems odd to include the normal distribution in the EFM; however, first recall the probability density function

$$f(y;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(y-\mu)^2}{2\sigma^2}). \quad (2\text{-}8)$$

Algebraic manipulation reveals that the normal distribution is indeed an EFM formulation.

$$EFM_{N(0,1)} = f_Y y(\theta,\phi) = \exp\{(\frac{(y\mu - \mu)^2/2)}{\sigma^2}) - \frac{1}{2}(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\} \quad (2\text{-}9)$$

Here, the natural (i.e., canonical) parameter is μ, and the scale parameter is σ^2.

These parameters need to be estimated. McCullagh and Nelder (1996) explain the estimation as follows: "In the case of generalized linear models, estimation proceeds by defining a measure of goodness of fit between the observed data and the fitted values that minimizes the goodness-of-fit criterion. We shall be concerned primarily with estimates obtained by maximizing the likelihood or log likelihood of the parameters for the data observed" (p. 23-24). This turns out to be the log of the EFM function.

$$\ell(\theta,\phi;y) = \log(f_Y(y;\theta,\phi)) \quad (2\text{-}10)$$

The natural and scale parameters are estimated by derivations revealing the mean function

$$E(Y) = \mu = b'(\theta), \quad (2\text{-}11)$$

and the variance function

$$\text{var}(Y) = b''(\theta)a(\phi). \quad (2\text{-}12)$$

Note that the mean function depends on only one parameter. However, as McCullagh and Nelder (1989) explain, " . . . the variance of Y is the product of two functions; one, $b''(\theta)$, depends on the [canonical] parameter (and hence on the mean) only and will be called the *variance function* [denoted $V(\mu)$], while the other [$a(\phi)$] is independent of θ and depends only on ϕ . . . The function $a(\phi)$ is commonly of the form $a(\phi) = \phi / w$" (p.29) and is commonly called the dispersion parameter. For the normal distribution, the natural parameter is the mean (μ); the variance function, $V(\mu)$; equals 1.0, and the dispersion parameter is σ^2.

The systematic component is simply the model for the predictors established as a linear combination and is denoted η. The link function, $g(\cdot)$, brings together the random component and the systematic component hence linking the function for the mean, μ, and the function for the systematic component, η, as $\eta = g(\mu)$. In other words, it specifies how the population mean of the outcome variable with a particular distribution is related to the predictors in the model. If $g(\mu)$ redundantly equals μ, then the population mean itself is related to the predictors. This is termed the identity link and is exactly the function used to link the mean of the normal distribution to its covariates.

The key advantage of GLM is that they are not restricted to one particular link function. Many other links are available. For example, one could specify the log link as $g(\mu) = \log(\mu)$ or the logit link as $g(\mu) = \log[\mu / (1-\mu)]$. However, each random component has one common 'canonical' link function that is best suited to the random component (McCullagh & Nelder. 1996). Alternatively, "Each potential probability distribution for the random component has one special function of the mean that is called its natural parameter" (Agresti, 1996, p.73). For example, a normal random component usually corresponds to an identity link, a Poisson

distributed random component usually corresponds to a log link, and a binomial distributed random component usually corresponds to a logit link. In sum, the canonical link and natural link are two equivalent terms for specifying the most suitable link connecting a particular distribution for the outcome variable with its linear systematic covariate function.

The Binomial Distribution and the Logit Link

Suppose we were interested in the differences between households with zero children under age 21 and households with one or more children over the age of 21. We could feasibly collapse all nonzero responses in the AEWR data into a value of one. Now, 71.58% of the values are zeros, and 28.42% of the values are ones. This distribution is obviously not normal. We have now introduced both a lower bound (zero), an upper bound (one), and an inherently nonnormal distribution; hence, a different random component and link can be specified to accommodate these constraints.

Variables that take on only one of two values are known as binary, or Bernoulli, variables, and the distribution of multiple independent trials for these variables is termed binomial. Bernoulli responses are modeled in terms of the probability (Pr) that the outcome variable (Y) is equal to either zero or one. The random component over multiple independent trials is thus a binomial distribution with parameters n for the number of trials and π for the probability that $Y = 1$.

$$Y \sim B(n, \pi) \tag{2-13}$$

The binomial distribution assumes that the responses are dichotomous, mutually exclusive, independent, and randomly selected (Agresti, 1996). Since the responses are discrete, the probability density function is termed the probability mass function and is defined as

$$f(k; n, p) = \frac{n!}{k!(n-k)!}[p^k (1-p)^{n-k}]. \tag{2-14}$$

This function gives the *p* probability of *k* ones (i.e., heads, hits, successes) over *n* trials. Rice (1995) clarifies, "Any particular sequence of *k* successes occurs with probability $[p^k(1-p)^{n-k}]$ from the multiplication principle [i.e., independent probabilities of realizations being multiplicative]. The total number of such sequences is $\frac{n!}{k!(n-k)!}$ [permutations], since there are $\frac{n!}{k!(n-k)!}$ ways to assign *k* successes to *n* trials" (p.36). The moment generating function is $\{1-\pi+\pi\exp(\xi)\}^n$, and the cumulant generating function is $n\log\{1-\pi+\pi\exp(\xi)\}$.

The log-likelihood function is virtually the same as the probability mass function. However, now we are determining the value of *p* as a function of *n* and *k* (rather than determining *k* as a function of *p* and *n*) while taking the log of this maximum at

$$\ell[f(p;n,k)] = \frac{n!}{k!(n-k)!}[p^k(1-p)^{n-k}]. \tag{2-15}$$

The estimates are obtained through derivations of the likelihood function as was previously discussed for the normal distribution. Just as the normal distribution population mean, μ, has the best maximum likelihood estimates of $\overline{X}$, the binomial distribution population probability, π, has the best maximum likelihood estimate of *k* divided by *n*, which is the proportion of ones, hits, or successes. This greatly reduces calculations when a quick estimate is needed and the random component is not linked to any predictors.

The binomial distribution eventually converges to a normal distribution. However, the speed of this convergence is primarily a function of skew,

$$Skew(y) = \frac{(1-2\pi)}{\sqrt{n\pi(1-\pi)}}, \tag{2-16}$$

with *p* = .50 yielding the fastest convergence (McCullagh and Nelder, 1996).

The link function should account for the binomial outcome variable. If linear predictors are used to predict a probability, then we have predicted values in an infinite range rather than constrained to be between zero and one. What is needed is a link function that will map a bounded zero-one probability onto this range of infinite values. The canonical link for the binomial distribution is the logit link.

$$\eta = g(\pi) = \log(\pi/(1-\pi)) \tag{2-17}$$

A logit is the natural log of an odds ratio or log[p / (1-p)]. An odds ratio is equal to a probability divided by one minus that probability. Hence, if the probability is .5, then the odds are .5 / (1-.5) = 1 meaning that the odds of a one are the same as the odds of a zero [i.e., the odds of success and failure are identical]. If the probability is .75, then the odds are .75 / (1-.75) = 3 meaning that a response of one is three times more likely than a response of zero. The reciprocal odds of 1/3 means that a response of one is three times

less likely than a response of zero, which is equivalent to stating that a response of zero is three times more likely than a response of one. When using the logit link to connect the binomial random distribution and the systematic component, the generalized linear model is

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta X_1 + ...\beta X_p . \tag{2-18}$$

A probability of .50, which is an odds of one, corresponds to a logit of zero. Odds favoring a response of one yield a positive logit, and odds favoring a response of zero yield a negative logit. Hence, the mapping is satisfied since the logit can be any real number (Agresti, 1996).

The regression parameters are slope-like in that they determine the relative rate of change of the curve. The exact rate of change depends on each probability with the best approximation

at that probability being $\beta\pi(1-\pi)$ with the steepest rate of change being at $\pi = 0.50$, which is where X = $-\alpha / \beta$ (Agresti & Finlay, 1997).

Since natural logs can be reversed through exponentiation and since odds can be converted to probabilities by dividing the odds by the sum of the odds and one, the fitted equation can be used to predict probabilities via

$$\pi = \frac{\exp(\alpha + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + [\exp(\alpha + \beta_1 X_1 + \cdots + \beta_k X_k)]}. \quad (2\text{-}19)$$

It is more common to interpret logistic regression coefficients by only exponentiating them. Then, the coefficient has a slope, rather than slope-like, interpretation; however, the relationship is multiplicative rather than additive. Specifically, the expected outcome is multiplied by exp(β) for each one-unit increase in X.

Evaluating Model Fit

The next step in interpreting generalized linear model results is to determine how well the estimated model fits the observed data, where fit is the degree of discrepancy between the observed and predicted values. McCullagh and Nelder (1989) explain, "In general the μs will not equal the *y*'s exactly, and the question then arises of how discrepant they are, because while a small discrepancy might be tolerable a large discrepancy is not" (p.33). The goodness of fit improves as the observed values and predicted values approach equality. For example, if a scatterplot reveals that all points fall on a straight line, then the predictive power of the regression equation would be perfect, and the subsequent fit would be perfect.

The comparison is usually performed through some statistical comparison of the observed outcome values and the predicted (i.e., fitted) outcome values. Rather than compare and summarize the actual observed and predicted values, it is common to gain summary information

by inspecting the log-likelihood value produced from the estimation procedure. Since the model parameters are estimated are from the data, perfect fit (i.e., observed-fitted = 0) is rare.[4] Hence, the goodness of the fit is measured to determine whether the difference is small enough to be tolerated.

There are many measures of model fit. Typically, the model is compared either to a null model in which the only parameter is the mean or a full model in which the number of parameters is equal to the sample size. "It is well-known that minus twice the LR statistic has a limiting chi-square distribution under the null hypothesis" (Vuong, 1989, p.308). McCullagh and Nelder (1989) equivalently state, "The discrepancy of a fit is proportional to twice the difference between the maximum log likelihood achievable and that achieved by the model under investigation" (p.33). This deviance statistic (G^2) is then considered to be asymptotically distributed chi-square with degrees of freedom equal to the number of parameters subtracted from the sample size.[5] A significant *p*-value indicates that the deviance is greater than what would be expected under a null hypothesis that the model with less parameters is adequate; hence, the observed model with an additional parameter or parameters is considered a significant improvement over the null model.

Another measure of model fit is Pearson's X^2; however, unlike G^2, it is not additive for nested models. Yet another measure of model fit is Akaike's Information Criterion (AIC), which penalizes the deviance for the number of parameters in the model.[6] The notion is that increasing

[4] Perfect fit is always obtained if the number of parameters and the sample size are identical (McCullagh & Nelder, 1989).

[5] The relationship is not always exact since sometimes the deviance is scaled and/or the likelihood is more difficult to estimate than in the simple logistic regression scenario presented here (McCullagh & Nelder, 1989).

[6] Other measures such as the Bayesian Information Criterion (BIC) penalize for both the number of parameters and the sample size.

the number of parameters will increase the log-likelihood regardless of the model and the data. Hence, the AIC penalizes the log-likelihood with regard to the number of parameters. There are two variations that provide further penalties. The Bayesian Information Criterion (BIC) penalizes for sample size; the Consistent Akaike Information Criterion (CAIC) penalizes even further by considering sample size and adding a small adjustment (Cameron & Trivedi, 1998). These indices can be compared to those of competing models; however, this must be done descriptively, not inferentially. The disadvantage is that the AIC can not be compared to a statistical distribution resulting in probabilities for significance testing; however, the advantage is that, as a descriptive statistic, it can be used to compare nonnested models.

The explanation thus far points to the fact that models can be compared to null, full, or other models. Statistical comparison is valid to the extent that one model is nested within the other, which is to say that both models share the same parameters, and one model has at least one parameter that is not included in the other. Alternatively, Clarke (2001) defines the models as follows: "Two models are nested if one model can be reduced to the other model by imposing a set of linear restrictions on the parameter vector . . . Two models are nonnested, either partially or strictly, if one model cannot be reduced to the other model by imposing a set of linear restrictions on the parameter vector" (p.727). The deviance for comparing two models is calculated as the difference in log likelihoods between the two models then multiplied by -2. This quantity is asymptotically distributed chi-square with degrees of freedom equal to the difference in parameters between the two models (Agresti, 1996). A significant *p*-value indicates that the deviance is greater than what would be expected under a null hypothesis of model equivalence; hence, the more complex model with an additional parameter or parameters is considered a significant improvement over the nested model.

The difference in log-likelihood statistics (i.e., deviance) can not be used to statistically test nonnested models. This is due to the fact that neither of the models can be considered the simple or more complex models with additional variables leading to a probabilistically higher log-likelihood. A *t*-test (or *F*-test) is a sensible alternative that eliminates concern for nesting. However, Monte Carlo simulations have demonstrated that, for model comparison tests, the *F*-test is lacking in sufficient power and can result in multicollinearity (Clarke, 2001).

The motivation for the AIC statistic is that, all else being equal, "the greater the number of coefficients, the greater the log-likelihoods" (Clarke, 2001, p.731). Hence, model fit becomes impacted by the number of variables in the model along with the effects of those variables. Hence, the AIC penalizes for the number of parameters. The formula is

$$AIC = -2(LL) + 2K \tag{2-20}$$

where *LL* is the log-likelihood estimate and *K* is number of parameters in the model including the intercept. Hence, now the log-likelihood is adjusted to accommodate simplicity and parsimony (Mazerolle, 2004).

In actuality, one could compare log-likelihoods between nonnested models. However, beyond the lack of parameter penalty, this technique might lead to the statistical hypothesis testing associated with log-likelihood statistics (i.e., test for the deviance approximated by χ^2).The AIC, on the other hand, should not be used in a formal statistical hypothesis test regardless of whether the model is nested or nonnested (Clarke, 2001). Generally, the researcher looks at several AIC indices and decides which model fits best based on a lower-is-better criterion. Mazerolle (2004) states, "The AIC is not a hypothesis test, does not have a *p*-value, and does not use notions of significance. Instead, the AIC focuses on the strength of evidence . . . and gives a measure of uncertainty for each model" (p.181).

Logistic modeling necessitated treating all nonzero numbers of children as a value of one. Depending on the research question, this may be a loss of valuable information (Slymen, Ayala, Arredondo, & Elder, 2006). Although sometimes it is necessary to model zero-inflated binomial data (Hall, 2000), specifying a binary distribution and logit link is not an ideal method for handling zero-inflated count data. The generalized linear model that specifies a binomial distribution and a logit link becomes more relevant when discussing the technique of splitting zero-inflated data into a model for the probability of zero separate from or combined with a model for the counts.

The Poisson Distribution and the Log Link

McCullagh and Nelder (1989), Lawal (2003), and Rice (1995) are the key references for the technical underpinnings for this model and distribution. The generalized linear 'Poisson' model is considered to be the benchmark model for count data (Cameron & Triverdi, 1998).[7] This is primarily attributed to the fact that the Poisson distribution has a nonnegative mean (Agresti, 1996). If y is a nonnegative random variable, the Poisson probability mass function is given by

$$f(k;\lambda) = \Pr(Y = k) = \frac{e^{-\lambda}\lambda^{y}}{k!}, \quad k = 0, 1, 2, \ldots \tag{2-21}$$

where λ is standard Poisson notation for the mean (μ) and k is the range of counts. Derivations by Rice (1995) show that the expected value of a random Poisson variable is λ; hence, "the parameter λ of the Poisson distribution can thus be interpreted as the average count" (p.113). Alternatively, lambda (λ) represents "the unobserved expected rate or occurrence of events . . . " (Zorn, 1996, p.1). The moment generating function is

[7] It is also commonly used to model event count data, which is "data composed of counts of the number of events occurring within a specific observation period . . . [taking the] form of non-negative integers (Zorn, 1996, p.1)".

$$E(e^{tX}) = e^{\lambda(\exp(t)-1)}. \quad (2\text{-}22)$$

The resulting cumulant generating function is $\lambda(\exp(t)-1)$, which, with a variance function of λ and a dispersion parameter equal to one, leads to mean and variance both being equal to λ, and the skew equal to one divided by the square root of λ. This equivalence of the mean and variance defined by a single parameter (Cameron & Triverdi, 1998; Agresti, 1996) is the result of a function that yields residuals that sum to zero (Jang, 2005); hence, the systematic portion of a Poisson GLM has no error term.

The Poisson distribution is a generalization of a sequence of binomial distributions. Rodríguez (2006) explained that "the Poisson distribution can be derived as a limiting form of the binomial distribution if you consider the distribution of the number of successes in a very larger number of Bernoulli trials with a small probability of success in each trial. Specifically, if $Y \sim B(n,\pi)$, then the distribution of Y as n $\rightarrow \infty$ and $\pi \rightarrow 0$ with $\mu = n\pi$ remaining fixed approaches a Poisson distribution with mean μ. Thus, the Poisson distribution provides an approximation to the binomial for the analyses of rare events, where π is small and n is large (p.3). Rice (1995) clarified, "The Poisson distribution can be derived as the limit of a binomial distribution as the number of trials, n, approaches infinity and the probability of success on each trial, p, approaches zero in such a way that $np = \lambda$" (p.43). Scheaffer (1995) and Rice (1995) have derived the generalization.

Further, just as a binomial distribution converges to a normal distribution given sufficient trials, the Poisson distribution converges to a normal distribution given a large mean.

The log-likelihood function for the Poisson distribution is

$$\ell(\lambda, y) = \sum y_i \log \lambda_i - \lambda_i, \quad (2\text{-}23)$$

with the maximum likelihood estimate of λ simply being the sample mean (Rice, 1995) and with the related deviance function being

$$Deviance(\lambda, y) = 2\sum\{y_i \log(y_i / \lambda_i) - (y_i - \lambda_i)\}. \quad (2\text{-}24)$$

McCullagh and Nelder (1989) state that the second term is often ignored. "Provided that the fitted model includes a constant term, or intercept, the sum over the units of the second term is identically zero, justifying its omission" (McCullagh & Nelder, 1989, p.34).

The systematic portion of the generalized linear model takes the form

$$\lambda_i = \exp(\mathbf{x}_i'\beta) = \exp(x_{1i}\beta_1)\exp(x_{2i}\beta_2)...\exp(x_{ki}\beta_k) \quad (2\text{-}25)$$

which is often equivalently expressed as

$$\log(\lambda_i) = \beta' X_i \quad (2\text{-}26)$$

with β derived by solving the equation

$$\sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i'\beta))\mathbf{x}_i = \mathbf{0} \quad (2\text{-}27)$$

by using iterative computations such as the Newton-Raphson. The canonical link for a generalized linear model with a Poisson random component specification is the log link (Stokes, Davis, & Koch, 1991).

$$\eta = \log(\lambda),\ Y \sim P(\lambda). \quad (2\text{-}28)$$

The Poisson distribution is not limited to count variates. Cameron and Triverdi (1998) explain that, although counts are usually in the purview of directly observable cardinal numbers, they may also arise through a latent process. In other words, ordinal rankings such as school course grades may be discretized as pseudocounts and assumed to have a Poisson distribution. Hence, the results of an analysis based on a Poisson distribution and the results using an ordinal analytic technique are often comparable.

As is almost always the case, it is common to identify other variables associated with the count variable (i.e., misspecification). However, the Poisson model has in interesting feature in that it assumes that there are no variables excluded from the model that are related to the count variable. In other words, there is no stochastic variation (i.e., no error term) (Cameron & Trivedi, 1998). Modifications must be made when one wishes to use a Poisson model with stochastic variation.

Iterative Estimation

Agresti (1996) clarifies the iterative estimation procedure. "The Newton-Raphson algorithm approximates the log-likelihood function in a neighborhood of the initial guess by a simpler polynomial function that has shape of a concave (mound-shaped) parabola. It has the same slope and curvature location of the maximum of this approximating polynomial. That location comprises the second guess for the ML estimates. One then approximates the log-likelihood function in a neighborhood of the second guess by another concave parabolic function, and the third guess is the location of its maximum. The successive approximations converge rapidly to the ML estimates, often within a few cycles" (p.94). The most common methods for estimating standard errors include Hessian maximum likelihood (MLH) (i.e., second partial derivative based) standard errors and maximum likelihood outer products (MLOP) (i.e., summed outer product of first derivative) estimation.

Interpretation of Coefficients

Standard ordinary linear squares regression lends an interpretation of β as the predicted additive change in the response variable per one-unit change in the predictor variable. However, as was the case with the binomial distribution, the interpretation differs when considering exponential distributions. For the Poisson distribution, "a one-unit increase in X has a multiplicative of $\exp(\beta)$ on the μ. The mean of Y at x+1 equals the mean of Y at x multiplied

by exp(β)" (Agresti, 1996, p.81).[8] Due to the inherent difficulty in interpretation, it is common to express in one of three alternative ways. First, the direction of the sign of β indicates a positive or negative 'effect' of the predictor on the count variable. Second, the fitted value can be calculated at the mean.[9] Third, some interpret the coefficient in terms of percent change; hence, if β=1.64, then as X increases to X+1, the predicted probability increases by 64% (Agresti, 1996).

Hypothesis testing

After conducting the analysis and estimating parameters, hypotheses can be tested in several ways as explained by Agresti (1996). One could test the hypothesis that β=0 using the traditional Wald *z*-test via $z = b / se_b$. Some programs provide Wald test results that are actually z^2; this is the Wald χ^2 statistic with one degree of freedom and appropriate only for a two-tailed test. A third method, the Score test, is "based on the behavior of the log-likelihood function at the null value for β=0" (Agresti, 1996, p.94) yielding a chi-square distributed statistic with one degree of freedom.

Overdispersion

In practice, the assumption of an equal mean and variance is the exception rather than the norm (McCullagh & Nelder, 1989). It is often the case that the sample variance is greater than or less than the observed sample mean with these two seldom being statistically equivalent (Cameron & Trivedi, 1998), especially for zero-inflated data (Welsh, Cunningham, Donnelly,

[8] This is similar to the interpretation for the binomial distribution with logit link; however, now the multiplicative effect is directly on μ rather than on the odds of μ.

[9] This can be particularly troublesome since that fitted value will only hold at the mean. It provides no valid inference for values greater than or less than the mean since the function is a curve with steepness that can vary drastically between separate values for the predictors.

and Lindenmayer, 1996). This condition is known as overdispersion[10] (underdispersion) and is a violation of a major tenet of the Poisson distribution that the conditional mean and conditional variance of the dependent variable are equal (i.e., equidispersion, nonstochasticity) (Jang, 2005; Zorn, 1996).[11] This assumption of equidispersion is the analog of the ordinary least squares regression assumption of homoscedasticity.

The overdispersion has been explained as heterogeneity that "has not been accounted for [that is] unobserved (i.e., the population consists of several subpopulations, in this case of Poisson type, but the subpopulation membership is *not* observed in the sample" (Böhning, Dietz, Shlattman, Mendonca, & Kirchner, 1999, p.195). The impact of violation is one of incorrect conclusions due to inaccurate *t*-statistics and standard errors (Cameron & Triverdi, 1998; Agresti, 1996). "The estimates of the coefficients can still be consistent using Poisson regression, but the standard errors can be biased and they will be too small" (Jewell & Hubbard, 2006, p.14). Alternatively, Slymen, Ayala, Arredondo, and Elder (2006) state that "Confidence intervals for regression estimates may be too narrow and tests of association may yield *p*-values that are too small" (p.2). The underlying mechanism for overdispersion is explained as unobserved heterogeneity in responses.

It is apparent that some modification to the variance to accommodate over-dispersion is ideal. Typically, maximum likelihood procedures are used to estimate parameters in the model. "The term *pseudo-* (or *quasi-*) *maximum likelihood estimation* is used to describe the situation in which the assumption of correct specification of the density is relaxed. Here the first moment [i.e., the mean] of the specified linear exponential family density is assumed to be correctly

[10] Overdispersion is sometimes referred to as extra-Poisson variation (Böhning, Dietz, Shlattman, Mendonca, & Kirchner, 1999).

[11] The under- or overdispersion may disappear when predictors are added to the model; however, this is likely not the case if the variance is more than twice the mean. (Cameron & Trivedi, 1989).

specified, while the second [i.e., the variance] and other moments are permitted to be incorrectly specified" (Cameron & Triverdi, 1998, p.19). Hence, the Poisson distribution as a baseline (Ridout, Demétrio, & Hinde, 1998) can be modified to accommodate overdispersion and underdispersion (Cameron & Triverdi, 1998).

Rice (1995) states that "gamma densities provide a fairly flexible class for modeling nonnegative random variables" (p. 52). One way to accommodate overdispersion is to consider the unobserved heterogeneity as a gamma distributed disturbance added to the Poisson distributed count data (Jang, 2005). In other words, an individual score may be distributed Poisson with a mean of λ but then this mean is "regarded as a random variable which we may suppose in the population to have a gamma distribution with mean μ and index μ/ϕ " (McCullagh & Nelder, 1989, p.199). This mixture leads to the negative binomial distribution. Given gamma function, Γ, and count *y*, the negative binomial probability mass function is

$$pr(Y = y; \mu, \phi) = \frac{\Gamma(y + \phi\mu)\phi^{\phi\mu}}{y!\Gamma(\phi\mu)(1+\phi)^{y+\phi\mu}} \tag{2-29}$$

Cameron and Trivedi (1998) provide a formulation where the negative binomial extends from the Poisson rather than being explained as a mixture distribution. Given a set of predictor variables, we can define

$$\omega_i = \mu_i + \alpha\mu_i^p \tag{2-30}$$

where α is scalar parameter to be specified or estimated, and *p* is a pre-specified power term. If the scalar parameter is set to zero then the resulting variance is equal to the mean, and the Poisson distribution holds. Hence, the Poisson model is nested within the negative binomial model.

The standard formulation for the negative binomial formulation of the function, sometimes called NB2 (Cameron & Trivedi, 1998), leads to a variance that is quadratic by setting *p* to 2

$$\omega_i = \mu_i + \alpha\mu_i^2 \tag{2-31}$$

This is the formulation seen in most textbooks (Rodríguez, 2006; Scheaffer, 1996). Its mean is the same as that of the Poisson distribution; however, its variance is derived from the gamma distribution (Cameron & Trivedi, 1998). Just as the Poisson distribution converges to a binomial distribution, the negative binomial distribution converges to a Poisson distribution (Jewell & Hubbard, 2006).

Poisson and Negative Binomial Models with Zero-Inflation

As elaborated upon previously, the zero-inflation problem is two-fold. First, the proportion of zeros is higher than expected given the specified population distribution shape resulting in an excess zeros problem. This can be descriptively determined by calculating the expected number of zeros as

$$E(fq(Y)) = fq(Y) * (\exp(-\overline{Y})) = n\,\mathrm{e}^{-\lambda} \tag{2-32}$$

For example, Zorn's example had a frequency of 4,052 with a $\lambda = 0.11$. The expected frequency of zeros would be

$$E(fq(Y)) = 4052 * (\exp(-0.109)) = 3,634\,, \tag{2-33}$$

which is less than the 3,882 zeros observed in the data. It turns out to be (3882/3634)*100 = 107% of the expectation (Zorn, 1996).

Second, the zeros can be a mixture of structural (i.e., true) zeros (Ridout, Demétrio, & Hinde, 1998) and sampled zeros reflecting the multiple sources of zeros problem. Shankar, Milton, and Mannering (1997) state that if "a two-state process is modeled as a single process . . .

if applying traditional Poisson and NB distributions, the estimated models will be inherently biased because there will be an over-representation of zero observations in the data, many of which do not follow the assumed distribution of [the] frequencies" (p.830)" Shankar, Milton, and Mannering (1997) note that the negative binomial model "can spuriously indicate overdispersion when the underlying process actually consists of a zero-altered splitting mechanism" (p.835-836). In sum, the sources of zeros arise from a dual generating process (i.e., structural and sampled) leading to two sources of unequal mean/variance dispersion (i.e., that due to unobserved heterogeneity of responses and that due to zero-inflation).

Most complex methods for analyzing zero-inflated count data model a mixture of two different distributions. The justification for splitting the distribution into two pieces is well-reasoned by Delucci and Bostrom (2004). "If it is deemed more reasonable to consider the zeros as indicators of cases without a problem, a more appropriate approach is to ask two questions: is there a difference in the proportion of subjects without the problem [i.e., structural true zeros], and, for those who have a problem [sampled false zeros], is there a difference in severity" (p. 1164).

Zorn (1996) refers to 'dual regime' models "wherein an observation experiences a first stage in which there is some probability that its count will move from a 'zero-only' state to one in which it may be something other than zero" (p.2). Typically, the dual-regime is composed of a *transition* stage based on a binomial distribution and an *events* stage based on some type of Poisson distribution.

There are many ways to model two-part distributions. For example, Mullahy (1986) and King (1989) proposed a Hurdle model in which the zeros are analyzed separately from the nonzeros. Lambert (1992) proposed a zero-inflated Poisson (ZIP) model in which different

proportions of zeros are analyzed separately and along with the nonzeros. Another early formulation (Heilborn, 1989) was the zero-altered Poisson (ZAP) model. "Arbitrary zeros are introduced by mixing point mass at 0 with a positive Poisson that assigns no mass to 0 rather than a standard Poisson (Lambert, 1992, p.1)". Mullahy (1986) presented a variation of the Hurdle model based on a geometric distribution[12] for use when specifying a Poisson distribution is not reasonable. Another possibility is to specify a log-gamma distribution for the event stage (Moulton, Curriero, & Barruso, 2002). Lambert (1989) presented a variation to the ZIP model known as ZIP(τ), which introduced a multiplicative constant to the event stage covariance matrix in order to account for the relationship between the two models. Gupta, Gupta, and Tripathi (1996) derived an adjusted generalized Poisson regression model for handling both zero-inflation and zero-deflation; however, accuracy was suggested to be contingent on the amount of inflation or deflation.[13]

It is also possible to formulate the model with different link functions. Lambert (1989) mentions the possibility of using the log-log link, complementary log-log link (Ridout, Demétrio, & Hinde, 1998), and additive log-log link while Lachenbruch (2002) mentions the lognormal and log-gamma distributions. Hall (2000) formulated a two-part model for zero-inflated binomial data. Gurmu (1997) describes a semi-parametric approach that avoids some distributional assumptions (Ridout, Demétrio, & Hinde, 1998).

There is some research on the extension of two-part models to accommodate random effects (Min & Agresti, 2005; Hall & Zhang, 2004; Hall, 2004; Hall, 2002; Olsen, 1999). Hall's (2000) model for zero-inflated binomial data permits both fixed and random effects. Dobbie and

[12] This distribution is an extension of the binomial distribution where the sequence in infinite. It is typically used in cases where the researcher is concerned with probability up to and including the first success (Rice, 1995).

[13] Min (2004) stated that the Hurdle model also has this feature.

Walsh (2001) permit correlated count data. Finally, Crepon and Duguet (1997) consider the cases where the variables are latent and correlated.

The evolution of the research to date has led to an emphasis on the standard Hurdle model and ZIP models (along with their negative binomial extensions) with a binary distribution for the transition stage and a Poisson distribution for the events stage and with fixed covariates. For both models, estimates are obtained from maximum likelihood procedures, although there has been some research on the use of generalized estimating equations (GEE) (Hall & Zhang, 2004; Dobbie & Welsh, 2001). The models are primarily distinguished by whether zeros are permitted in the event stage. In other words, their differences are a reflection of the researcher's notions about the potentially multiple sources of zeros in the data and their relationship to excess zeros.

They also differ in terms of the transition stage cumulative probability function. To be clarified in the sections that follow, Zorn (1996) summarizes the differences as follows: " . . . the hurdle model has asymmetric hurdle probability while in the ZIP specification p_i is symmetrical. Also, the hurdle model does not permit zero values to occur once the 'hurdle' has been crossed, while in the ZIP model zeros may occur in either stage" (p.4). Choosing between the models is a matter of validity; hence, the choice rests on substantive ground as well as statistical considerations. As Martin, Brendan, Wintle, Rhodes, Kuhnert, Field, Low-Choy, Tyre, and Possingham (2005) note, "it is imperative that the source of zero observations be considered and modeled accordingly, or we risk making incorrect inferences . . . " (p.1243-1244).

The Hurdle model

The Hurdle model was developed separately by Mullahy (1986) in economics and King (1989) in political science, although the term itself was most likely coined by Cragg (1971). [14]

[14] Cragg (1971) proposed the basic two-step process in which the probability of occurrence is modeled separately from frequencies of occurrence.

Welsh, Cunningham, Donnelly, and Lindenmayer (1996) refer to it as a 'conditional Poisson model'. "The idea underlying the hurdle formulations is that a binomial probability model governs the binary outcome whether a count variate has a zero or a positive realization [i.e., a transition stage]. If the realization is positive the 'hurdle' is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero[15] count data model [i.e., events stage]" (Mullahy, 1986, p.345) such as a truncated Poisson or truncated negative binomial distribution (Min & Agresti, 2005). In other words, one distribution addresses the zeros while another distribution addresses the positive nonzero counts. For example, for grade-retention data, there would be a model for schools with no dropouts and a model for school with at least one dropout.

It is a "finite mixture generated by combining the zeros generated by one density with the zeros and positives generated by a second zero-truncated density separately . . . " (Cameron & Trivedi, 1998, p.124). Log-likelihood values are estimated separately for each density. A key feature of the transition model is asymmetry in that the probability of crossing the hurdle increases more quickly as the covariates increase than it decreases as the covariates decrease. The function is then "asymmetrical" (King, 1989) leading to validity concerns supporting or refuting the substantive theory underlying the model.

The two-part distribution of the dependent variable is given first by the transition stage g_1 probability mass function

$$P(Y_i = 0) = g_1(0)\ ^{16} \tag{2-34}$$

[15] The lower bound is then one.

[16] This is an alternative representation of the aforementioned $\Pr(Y = 0) = 1 - \pi$.

modeling whether the response crosses the hurdle of zero. Assuming a Poisson distribution and log link, Zorn (1996) expands the cumulative distribution function to include covariates as

$$p = 1 - \exp[-\exp(X_{0i}^{'}\beta_0)] \tag{2-35}$$

The basic model for the event stage is then the probability for a nonzero realization multiplied by the probability for the counts. [17]

$$P(Y_i = j) = (1 - g_1(0))\frac{g_2(j)}{1 - g_2(0)} \ldots j = 1, 2, \ldots \tag{2-36}$$

Greene (1994) notates the models with a binomial distribution and logit link for the transition stage and a Poisson distribution with a log link for the event stage as[18]

$$\text{Transition: } \Pr(y_i = 0) = p \tag{2-37}$$

$$\text{Event: } \Pr(y_i = k) = (\frac{1 - p}{1 - e^{-\lambda_i}})(\frac{e^{-\lambda_i}\lambda_i^{k}}{k!}), k = 1, 2, \ldots \tag{2-38}$$

Here, p is the probability of a count of zero while λ is the truncated Poisson mean for the counts greater than zero. The generalized linear models as a function of covariates is then

$$\text{Transition Stage: } \log(\frac{p_i}{1 - p_i}) = \mathbf{x}_{1i}^{'}\mathrm{B}_1 \tag{2-39}$$

$$\text{Event Stage: } \log(\lambda_i) = \mathbf{x}_{2i}^{'}\mathrm{B}_2 \tag{2-40}$$

The two vectors of parameters are estimated jointly.

$$\Lambda_1 = [\prod\{\exp[-\exp(X\beta)]\}\prod\{1 - \exp[-\exp(X\beta)]\} \tag{2-41}$$

$$\Lambda_2 = [\prod\exp(yX\beta)/(\{\exp[\exp(X\beta)] - 1\}y!)] \tag{2-42}$$

An alternative notation provided by Min (2003) is

---

[17] The second term is an alternative representation of the aforementioned $f(k;\lambda) = \Pr(Y = k) = \frac{e^{-\lambda}\lambda^{y}}{k!}$.

[18] Note that p is used to representative the probability of a zero rather than the conventional representation of a probability of one.

$$\ell_1(\beta) = \sum_{y_i=0}[\log P_1(y_i = 0; \beta_1, x_{1i})] + \sum_{y_i>0}[\log(1 - P_1(y_i = 0; \beta_1, x_{1i}))]$$

$$= \sum_{y_i=0} x_{1i}, \beta_1 - \sum_{i=1}^{n} \log(1 + e^{x_{1i}'\beta_1}) \quad (2\text{-}43)$$

$$\ell_2(\beta) = \sum_{y_i>0}[y_i x_{2i}'\beta_2 - e^{x_{2i}'\beta_2} - \log(1 - e^{-e^{x_{2i}'\beta_2}})] - \sum_{y_i>0} \log(y_i!) \quad (2\text{-}44)$$

Since the two models are functionally independent, the likelihood functions can be maximized separately (Min & Agresti, 2005; Min, 2003, Cameron & Trivedi, 1998; Mullahy, 1986).

$$\Lambda_{hurdle} = \log(\Lambda_1) + \log(\Lambda_2) \quad (2\text{-}45)$$

$$\ell(\beta_1, \beta_2) = \ell(\beta_1) + \ell(\beta_2) \quad (2\text{-}46)$$

This is because "the large sample covariance between the two sets of parameters is zero so the joint covariance matrices can be obtained from the separate fits" (Welsh, Cunningham, Donnelly, & Lindermayer, 1996, p.300). Solving the likelihood equations uses either the Newton-Raphson algorithm or the Fisher scoring algorithm, both giving equivalent results (Min, 2003). The Poisson model is nested within the Hurdle model (Zorn, 1996). Hence, fit of these two models can be compared statistically.[19]

The Negative Binomial Hurdle model

In the case of zero-inflated data, it is possible to have two sources of overdispersion. The variance can be greater than the mean due to the preponderance of zeros. However, there is now the possibility of unobserved heterogeneity in the event stage (Mazerolle, 2004; Min & Agresti, 2004). The former scenario has been referred to as zero-driven overdispersion (Zorn, 1996); the latter is Poisson overdispersion. Just as was the case with the Poisson model, it is possible to nest the Hurdle model within a more general negative binomial framework. Further, the negative binomial Poisson model is nested within the negative binomial Hurdle model. Hence, the fit of a)

[19] Zhou and Tu (1999) developed likelihood ratio for count data with zeros; however, it was not generalized to any particular zero-inflation model.

the Poisson model and the Hurdle model, b) the negative binomial Poisson model and the negative binomial Hurdle model, and c) the Hurdle model and the negative binomial Hurdle model can be compared using statistical tests.[20] Estimation is typically performed by solving the maximum likelihood equations using the Newton-Raphson algorithm.

The Zero-Inflated Poisson (ZIP) model

The Zero-Inflated Poisson, or ZIP, model is another model that one can use when the zeros in a dataset are argued to be caused by both chance and systematic factors (Min & Agresti, 2005). The transition stage addresses zero-inflation while the event stage addresses unobserved heterogeneity of responses including zeros (Jang, 2005). Welsh, Cunningham, Donnelly, and Lindenmayer (1996) refer to it as a mixture model.

This two-part model, developed by Lambert (1992) permits zeros to occur in both the transition stage and event stage (Cameron & Trivedi, 1998); "crossing the 'hurdle' in the ZIP model does not guarantee a positive realization of Y" (Zorn, 1996, p.4). Further, the probability function in the transition stage is now symmetrical (Zorn, 1996). Lachenbruch (2002) explains that "ZIP regression inflates the number of zeros by mixing point mass at 0 with a Poisson distribution" (p. 12). Zorn (1996, p.4) clarifies the distinction between the ZIP and Hurdle models as follows:

> As a special case of the general model, the ZIP regression is thus seen to make substantially different assumptions about the nature of the data generating process than the hurdle model. Whether parameterized as a logit or a probit, the probability exiting the zero-only stage is assumed to follow a symmetric cumulative distribution. Likewise, even those cases which make the transition to the events stage may nevertheless have zero counts; crossing the "hurdle" in the ZIP model does not guarantee a positive realization of Y . . . The sole difference in assumptions here is that the hurdle model's count distribution is assumed to be truncated at zero whereas the ZIP specification count data may take on zero

[20] It is not the case that the negative binomial Poisson model is nested within the Hurdle model; hence, one can not statistically compare all four models collectively.

values in the event stage.

Another difference is that, unless the ZIP model is overly parameterized, only the Hurdle model can handle zero deflation (Min & Agresti, 2005).

Compared to the Hurdle model, the equations for the event stage are very similar. The exception is that (1-p_i) is divided by $(1-e^{-\lambda_i})$ in the Hurdle model before being multiplied by the remaining elements of the equation. However, the transition stage equations are strikingly different. For the Hurdle model, the equation is $\Pr(y_i = 0) = p$; the ZIP model includes the addition of the probability of a nonzero multiplied by the exponentiated Poisson mean. This is the mathematical characteristic that distinguishes the Hurdle model's exclusion of zeros in the event stage and the ZIP model's potential inclusion of zeros in the event stage. Rather than model the probability of a zero in the transition stage, the ZIP also models the probability that the counts have a Poisson distribution hence permitting zeros from both a perfect state and a Poisson state (Hur, 1999). Given this, λ parameterizes the mean of this Poisson distribution (Welsh, Cunningham, Donnelly, & Lindenmayer, 1996). When adding covariates, the Hurdle and ZIP generalized linear model appear the same.

$$\text{Transition Stage: } \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_{1i}'\mathbf{B}_1 \tag{2-47}$$

$$\text{Event Stage: } \log(\lambda_i) = \mathbf{x}_{2i}'\mathbf{B}_2 \tag{2-48}$$

Unlike the Hurdle model, the ZIP model likelihood function can not be maximized separately for the transition and event stage. Hence, the Hurdle model "has the attractive advantage of an orthogonal parameterization which makes it simpler to fit and interpret than the mixture model" (Welsh, Cunningham, Donnelly, & Lindenmayer, 1996) with the disadvantage of asymmetrical transition to counts. The likelihood function derived by Lambert (1992) is

$$L = \sum_{y_{i=0}} \log(e^{G_i\gamma} + \exp(-e^{B_i\beta})) + \sum_{y_i>0} (y_i B_i \beta - e^{B_i\beta}) - \sum_{i=1}^{n} \log(1 + e^{G_i\gamma}) - \sum_{y_i>0} \log(y_i!) \qquad (2\text{-}49)$$

where βB is the vector of coefficients and matrix of scores for the event stage and γG is the vector of coefficients and matrix of scores for the transition stage, and where iterations are based on the EM or Newton-Raphson algorithms (Min, 2003; Lambert, 1992)[21].

Strictly speaking, the Poisson model is not nested within the ZIP model; therefore, it would not be wise to conduct a formal model fit test (Zorn, 1996; Greene, 1994). However, it is interesting to note that the log-likelihood of 10,607 is slightly lower than that produced by the negative binomial Hurdle model. This is in line with Greene's (1994) observation when using the Vuong statistic as an alternative for testing nonnested models. "For present purposes, the important question is whether the ZIP models . . . provide any improvement over the basic negative binomial . . . The log-likelihood functions are uniformly higher, but as noted earlier, since the models are not nested, these are not directly comparable. The Vuong statistic, however, is consistent with the observation" (Greene, 1994, p.26).

The Negative Binomial Zero-Inflated Poisson model

The ZIP model can be extended to the negative binomial model just as the Poisson was extended to the negative binomial and as the Hurdle was extended to the Hurdle negative binomial. This may be necessary as Min (2003) explains that "Sometimes such simple models for overdispersion are themselves inadequate. For instance, the data might be bimodal, with a clump at zero and a separate hump around some considerably higher value. This might happen for variables for which a certain fraction follows some distribution have positive probability of a zero outcome" (p.13). He further explains, "The equality of the mean and variance assumed by

[21] Alternatively, Martin, Brendan, Wintle, Rhodes, Kuhnert, Field, Low-Choy, Tyre, and Possingham (2005), compared relative means and credible intervals estimate from a bootstrap procedure.

the ZIP model . . . is often not realistic. Zero-inflated negative binomial models would likely often be more appropriate than ZIP models" (p.15). Unfortunately, current iterative techniques lead to a greater risk of nonconvergence than when using the other two-part models (Fanoye & Singh, 2006; Lambert, 1992).

The key assumption in accepting these results over those produced from the Hurdle models is that there are some zeros that belong with the counts representing no household members under the age of 21 for reasons other than their never having such household members at all. Although it is not valid to statistically compare the fit of the ZIP model with the Hurdle and Poisson models, it is reasonable to test the fit of the ZIP model within the negative binomial ZIP model.

In sum, the choice between Hurdle models and ZIP models is ultimately guided by the assumptions one makes about the data generating process. Min (2004) states, "The zero-inflated models are more natural when it is reasonable to think of the population as a mixture, with one set of subjects that necessarily has a 0 response. However, they are more complex to fit, as the model components must be fitted simultaneously. By contrast, one can separately fit the two components in the hurdle model. The hurdle model is also suitable for modeling data with fewer zeros than would be expected under standard distributional assumptions" (p.20).

Model Comparison Testing for Zero-Inflated Data

The Poisson model is nested within the negative binomial Poisson differing only by the dispersion parameter. In fact, the two models are equivalent when one uses the negative binomial model and restricts the dispersion parameter to 1.0 achieving Poisson equidispersion (Cameron & Trivedi, 1998). Likewise, the Hurdle model is nested within the negative binomial Hurdle model, and the ZIP model is nested within the negative binomial ZIP model. However, since each involves estimation of a transition stage and event stage model, the nesting rule applies to

both equations. In other words, if the event stage contains 3 parameters and the transition stage 4 parameters for the nested model, then the more complex model must contain at least these same 3 parameters in the event stage and at least the same 4 parameters in the transition stage.

According to Zorn (1996), the Poisson model is nested within the Hurdle model, and the negative binomial Poisson model is nested within the negative binomial Hurdle model. This is reasonable given that the Hurdle models are estimating the Poisson models in the event stage and that these likelihood statistics are independent of those produced in the transition stage. The ZIP models, on the other hand, are not estimated in this manner; it is not reasonable to assume that the Poisson models are nested within the ZIP Models (Greene, 1994). This leads to the hierarchy of permissible model-testing displayed in Table 2-1. Other models can be compared descriptively using the aforementioned Akaike's Information Criterion (AIC). The remainder of the dissertation will use the deviance statistic for model comparison inference and the AIC for model comparison description.

Review of Research Pertaining to and Using Zero-Inflated Count Data

Hurdle Model

Statistical

As previously discussed, Mullahy (1986) presented the underlying statistical foundations for the Hurdle model. He specified it in terms of extending from both the Poisson and geometric distributions. He also presented an extension to the Hurdle model that he termed the "with-zeros" (WZ) model. The WZ model adjusts for zero-inflation by augmenting or reducing the probability by an additive constant (rather than having them specified by the parent distribution). Subsequent research has focused on the Hurdle model rather than the WZ model; this is most likely due to the fact that the WZ turns out to be a special case that collapses into a Hurdle model in some specifications, and estimates are often similar between the models (Mullahy, 1986). Finally,

Mullahy presented tests for specifications including a technically complex information matrix test as well as the Score test typically provided in software output. King's (1989) key contribution to the Hurdle model was a Monte Carlo study confirming that the counts can be viewed as arising from a Poisson process (Civetinni & Hines, 2005).

Min and Agresti (2004) conducted a simulation to compare a zero-inflation and a zero-deflation condition. They found that the estimates were reasonable for the ZIP model under zero-inflation. However, the coefficient and standard error for the event stage were both very large under zero-deflation. They explain that logit models simply can not accommodate too few zeros. "The zero-inflated model is only suitable for zero-inflation problems. However, the hurdle model is also suitable for modeling data with fewer zeros than would be expected under standard distributional assumptions. In fact, when a data set is zero-deflated at some levels of the covariates, the zero-inflation model may fail" (Min & Agresti, 2004, p.5).

In contrast, Mullahy (1986) stated that, "a particularly interesting feature of the modified count data specifications considered here [i.e., Hurdle model] is that they provide a natural means for modeling overdispersion or underdispersion of the data. Specifically, overdispersion and underdispersion are viewed as arising from a misspecification of the maintained parent [data generating process] in which the relative probabilities of zero and non-zero (positive) realizations implied by the parent distribution are not supported by the data" (Mullahy, 1986, p.342).

Applications

Mullahy (1986) researched daily consumption of coffee, tea, and milk where each is a count variable. Covariates included age, years of completed schooling, family income, sex, race, and marital status. For the Hurdle model, the test statistics [*p*-values] for the coffee-model were substantially smaller than those for tea-model and milk-model. Mullahy argued that this is not surprising given the ratio of estimates and standard errors. However, closer inspection of the data

reveals that it is the coffee variable that has the lowest proportion of zeros at only 26.26%. It is possible that the Hurdle model was unable to adequately account for the additional overdispersion witnessed in the other two models (61.63% and 40.37% zeros). In other words, a two-part model may be a necessary but not sufficient condition for handling overdispersion in zero-inflated count models, and negative binomial formulations may be increasingly necessary as the proportion of zeros increases.

King (1989) applied the Hurdle model to data for the relationship between the number of nations entering war in a period of time as a function of those in formal international alliances. The hurdle model was formulated based on the premises of Mullahy (1986) and justified due to there being some countries who will not go to war and others who will not at first but will later be 'dragged in' by alliances. Hence, this is a classic example of the justification for true zeros and event-driven zeros. This was demonstrated statistically by comparing Hurdle and Poisson results. The Poisson alliance coefficient of .007 was significant, the Hurdle model event-stage coefficient of .007 was significant, and the Hurdle model transition stage coefficient of .001 was not significant. Hence, the Poisson interpretation would be that increased alliances lead to increased war frequency. However, the Hurdle results clarify that this is only true after the onset of war (i.e., the hurdle has been crossed). Further statistical evidence supported the Hurdle model based on the likelihood-ratio model comparison test.

Zero-Inflated Poisson Model

Statistical

As previously discussed, Lambert (1989) presented the original formulation for the ZIP model and its ZIP(τ) formulation. She also presented the models' extension from the Poisson and negative binomial as well as the derivation of the maximum likelihood (EM) estimates. She ran several simulations to test the adequacy of the model. The first simulation varied sample size,

with one covariate taking on a fixed coefficient and a fixed variance in both parts of the model. The result was an average of 50% zeros in the transition stage and 23% zeros in the event stage. The results suggest that the ZIP model consistently converges at n=25 when using EM and at n=100 when using the Newton-Raphson algorithm. An examination of confidence intervals revealed that the normal-theory intervals are not reliable at n=100; however, almost all simulated likelihood-ratio confidence intervals contained the true mean even at n=25. "To summarize, these simulations with one covariate for both λ and p are encouraging. The ZIP and ZIP(τ) regressions were not difficult to compute, and as long as inference was applied only when the observed information matrix was nonsingular, estimated coefficients, standard errors based on observed information, and estimated properties of Y could be trusted" (Lambert, 1992, p.7).

Warton (2005) compared 20 datasets of varying sample sizes, proportions of zeros, and factors/levels. The ordinary least squares version included the addition of one to all counts before taking the logarithm. The other models not accommodating zero-inflations were the Poisson and four formulations of the negative binomial Poisson (including the aforementioned quasi-Poisson where the variance is set to $\varphi\mu$). The zero-inflated models included the ZIP model and the negative binomial ZIP model. The Akaike Information Criterion (AIC) values were calculated for a total of 1,672 variables averaged over datasets and rescaled to a minimum AIC of zero for each dataset.

As expected, when overdispersion was present, the negative binomial formulations outperformed the models without these formulations. However, when overdispersion was not present, the reverse was true for 53% of the variables. This suggests that the level of skew in the model interacts with zero-inflation when measuring model adequacy.

When the proportion of zeros is very small, the distribution looks more like a Poisson distribution truncated at zero. In other words, it shares features modeled by the event stage of a Hurdle model. This led to estimation problems in which the negative binomial model ZIP model rarely converged. When it did converge, its fit was better than that of the ZIP model for only 11% of the datasets. Possibly, as previously discussed, the zero-inflated Hurdle would converge more often since it can handle both zero-inflation and zero-deflation (Min & Agresti, 2005; Min & Agresti, 2004).

A very interesting finding pertained to the transformed OLS fit indices. "Although transformed least squares was not the best fitting model for data, it fitted the data reasonably well. Surprisingly, transformed least squares appeared to fit data about as well as the zero-inflated negative binomial model . . . The AIC for transformed least squares was not as small as for the negative binomial model overall, although it was smaller for 20 per cent of the variables considered here" (Warton, 2005, p.283).

Averaging over all datasets, the AIC was lowest for all negative binomial models followed by a close tie between the transformed OLS model and the negative binomial ZIP model, which was followed by the ZIP model. All models were a drastic improvement over the standard Poisson model. The implications are that, although the Poisson is rarely adequate when the data is not equidispersed and/or is inflated or deflated, an intuitive climb up the ladder of models may not be reasonable. There were features of these datasets including varying degrees of zero-inflation and overall distributions that warrant further investigation toward appropriate model selection. "If one were to fit a zero-inflated model, it would be advisable to present quantitative evidence that the zero-inflation term was required. Based on the present results, it is likely that a term for extra zeros is not needed, and a simpler model will usually suffice . . .

special techniques are not generally necessary to account for the high frequency of zeros. The negative binomial was found to be a good model for the number of zeros in counted abundance datasets, suggesting that a good approach to analyzing such data will often be to use negative binomial log-linear models" (Warton, 2005, p.287-288).

In regard to their problems with zero-inflated negative binomial convergence, Fanoye and Singh (2006) developed an extension that improves convergence termed the zero-inflated generalized Poisson regression (ZIGP) model. Their recent research revealed convergence in less than 20 iterations for all trials. However parameter estimates and standard errors were often very different than those produced by the ZIP model. They conclude, "Even though the ZIGP regression model is a good competitor of ZINB regression model, we do not know under what conditions, if any, which one will be better. The only observation we have in this regard at this time is that in all of the datasets fitted to both models, we successfully fitted the ZIGP regression model to all datasets. However, in a few cases, the iterative technique to estimate the parameters of ZINB regression model did not converge" (p.128).

Greene (1994) used the Vuong statistic when comparing the Poisson, negative binomial Poisson, and ZIP models. It was noted that the rank order for the Vuong statistics and the log-likelihood estimates were in alignment. The conclusion suggested future research using the Vuong statistic. "The use of Vuong's statistic to test the specification seems not to have appeared in the recent literature . . . We conjecture that the Vuong testing procedure offers some real potential for testing the distributional assumption in the discrete data context. In the cases examined, it appears to perform well and in line with expectations" (Greene, 1994, p.30).

Shankar, Milton, and Mannering (1997) used the Vuong statistic to decide between the negative binomial, ZIP, and negative binomial ZIP model for traffic accident data. They clarify

the interpretation of the statistic stating, "A value >1.96 (the 95% confidence level of the *t*-test) for *V* favors the ZINB while a value < -1.96 favors the parent-NB (values in between 1.96 an -1.96 mean that the test is indecisive) . . . This test can also be applied for the ZIP(τ) and ZIP cases" (p.831).

Civettini and Hines (2005) explored the effects of misspecification on negative binomial ZIP models. This included misspecification by leaving a variable out of the event stage that was present in the event stage and misspecification by shifting a variable from the transition stage to the event stage.

Applications

Lambert (1992), in formulating the ZIP model, applied it to the analysis of defects in manufacturing. In terms of improperly soldered leads, 81% of circuit boards had zero defects relative to the 71% to be expected under a Poisson distribution linked to a model with a three-way interaction. This most complicated model had a log-likelihood of -638.20. This dropped to -511.2 for the ZIP model. Although comparing to a different combination of covariates, the negative binomial Poisson model fit better than the Poisson model but not as well as the ZIP model.

Greene (1994) used credit-reporting data to investigate differences between the Poisson, negative binomial Poisson, ZIP, negative binomial ZIP, as well as some of their aforementioned variants and the specification of a probit link rather than the logit link. The data consisted of 1,023 people who had been approved for credit cards. The count variable of concern was the number of major derogatory reports (MDR), which is the number of payment delinquencies in the past 60 days.

For this sample, 89.4% had zero MDR. Given a mean of 0.13, this frequency of 804 is nearly double the 418 we might expect in a Poisson distribution. The skew of 4.018 is reduced to

2.77 when ignoring the zeros while the mean increases to 1.22. As expected, the negative binomial Poisson resulted in improved fit (based on the Vuong test statistic), increased standard errors and different parameter estimates. The ZIP model resulted in slightly worse fit than the negative binomial Poisson while remaining much better compared to the Poisson model. If all of the overdispersion was due to unobserved response heterogeneity then the results should be similar for the negative binomial ZIP model. However, this model produced the best fit of all.

It is interesting to note that, again, the standard errors increase while the parameter estimates are different relative to the ZIP model. In fact, of the 6 parameters, 4 estimates decreased, 2 increased, and 1 switched in sign. Hence, there are two implications. First, the negative binomial ZIP model was necessary to accommodate two sources of overdispersion to adjust standard errors. Second, ignoring the negative binomial formulations would have led to nonsensical parameter estimates driven by a sample mean of 0.13.

Böhning, Dietz, Schlattmann, Mendonça, and Kirchner (1999) compared pre- and post-intervention scores on the decayed, missing, and filled teeth index (DMFT) for 797 children in one of six randomly assigned treatment conditions. The results were not exhaustive; however, the log-likelihood did decrease from -1473.20 to -1410.27 when going from the Poisson model to the ZIP model. This study was somewhat unique in that all the covariates (sex, ethnicity, and condition) were categorical, and that the conditions were dummy-coded represented as five parameters. Also, this data had features that might suggest that zero-inflated models weren't necessary. For pre-intervention, the proportion of zeros was 21.58%, which increased to only 28.99% at post-intervention. The means, with the zeros in the data, were 3.24 and 1.85, respectively. Ignoring the zeros changed these means to 4.13 and 2.61, respectively. The skew, with the zeros in the data, was 0.20 and 0.65, respectively. Ignoring zeros changed the skew to

0.08 and 0.63, respectively. In other words, many features of the data were consistent with what would be expected of a Poisson, and possibly normal, distribution. Nonetheless, with these means and frequencies, the Poisson distribution suggests overdispersion with 31 permissible zeros for the pre-intervention and 125 permissible for the post-intervention whereas the data revealed 173 zeros and 232 zeros, respectively. It then becomes a matter of whether the overdispersion was due to the proportion of zeros in each condition or unobserved heterogeneity in the event stage. The negative binomial ZIP model was not used to analyze this data.

Xie, He, and Goh (2001) analyzed the number of computer hard disk read-write errors. Approximately 87% of the 208 cases were zeros. Given that the Poisson mean was 8.64, the authors noted that the ZIP model is to be preferred over the Poisson model. However, this mean is due to several values between 1 and 5, a few between 6 and 15, and 2 values of 75. These latter two values appear to be so far from the others that they should have been treated as outliers and addressed in some other manner.

Jang (2005) analyzed the number of non-home based trips per day from 4,416 households in Jeonju City, Korea. The provided bar graph suggested that approximately 45% of the cases were zeros. The Vuong statistic (Vuong, 1989) was used for model selection given that the Poisson is not nested within the ZIP or negative binomial ZIP models.

The purpose of the article by Delucchi and Bostrom (2004) was to provide a brief introduction to many possible methods for handling zero-inflation including standard *t*-tests, bootstrapping,[22] and nonparameteric methods. In doing so, they provided results from a study involving 179 patients with opioid dependence assigned to either a methadone-maintenance or methadone-assisted-detoxification treatment. Five out of seven ways to segment the sample

[22] See Jung, Jhun, and Lee (2005) for bootstrap procedures and simulation results for Type I and Type II errors.

resulted in zero-inflation ranging from 17% to approximately 66% zeros. The only two-part model to be used was the ZIP model. The table of results revealed that, in terms of *p*-values, the ZIP model performs either very similarly or very differently from the Pearson χ^2 test for the proportion of zero values, the Mann-Whitney-Wilcoxon test of nonzero values, and/or the Mann-Whitney-Wilcoxon test of difference in mean scores between treatment groups. It is possible that these tests become more similar as the proportion of zeros declines but such conclusions are based purely on the table of *p*-values.

Desouhant, Debouzie, and Menu (1998) researched the frequency of immature weevils in chestnuts. One tree was measured over 16 years, another was measured over 11 years, and three trees were measured on 1 year. The means ranged from .06 to .63. "None of the 30 distributions fits a Poisson, χ^2 values being always very significant . . . The ZIP distribution fits 25 out of 31 cases . . . The NB distribution fits 20 out of the 31" (Desouhant, Debouzie, & Menu, 1998, p.384). This led to the conclusion that researchers should consider both true zeros and overdispersion (i.e., trees noted as 'contagious' and trees varying in random oviposition behavior).

Shankar, Milton, and Mannering (1997) analyzed a 2-year summary of traffic accident frequencies. For principal arterials, they chose a negative binomial model with data ranging from 0 to 84 ($M = 0.294$, $SD = 1.09$). For minor arterials, they chose the negative binomial ZIP model for data ranging from 0 to 7 ($M = 0.09$, $SD = 0.346$). For collector arterials, they chose the ZIP model for data ranging from 0 to 6 ($M = 0.61$, $SD = 0.279$). Model selection was based on the Vuong statistic. For example, they state, "As suspected previously, inherent overdispersion in the data is due to the parent NB process and this was validated when the [negative binomial ZIP]

specification failed to provide a statistically better fit (the Vuong statistic <1.96, which corresponds to the 95% confidence limit of the *t*-test" (p.833).

Slymen, Ayala, Arredondo, and Elder (2006) analyzed percent calories from fat and number of days of vigorous physical activity from 357 females participating in a baseline condition and one of three treatment conditions. Covariates included employment status, education, martial status, cigarette smoking, and self-reported health. The zero-inflation was 294 out of 357 (82.4%). They compared models using likelihood ratio tests between the Poisson and negative binomial Poisson and likewise between the ZIP and negative binomial ZIP. The AIC's were inspected to compare the Poisson and ZIP models.

Not surprisingly, the negative binomial model fit better than the Poisson model. However, the ZIP model did not fit better or worse than the negative binomial ZIP, and the parameter estimates and standard errors were nearly identical. This suggests almost no overdispersion in the data. Indeed, the nonzero percentages were as follows: 1 = 2.8%, 2 = 3.4%, 3 = 4.8%, 5 = 2.0%, 6 = 0.0%, and 7 = 2.0%. This suggests strong equidispersion leaning toward a uniform nonzero distribution. The AIC's for both models were also nearly equal although both being considerably smaller than the AIC for the Poisson model and somewhat smaller than the AIC for the negative binomial Poisson model. Based on a 'smaller-is-better' heuristic, the authors favored the ZIP model with an AIC of 562.5 over the zero-inflated ZIP model with an AIC of 565.

ZIP and Hurdle Model-Comparisons

The purpose of this section is to present a breadth of literature in which both the Hurdle and ZIP models were either compared statistically and/or used to analyze real data. This also includes extensions such as the negative binomial and tau formulations (e.g., ZIP(τ)). Some authors presented alternatives that seem to depart from the ZIP and Hurdle models too drastically to be within the scope of this dissertation. For example, Lachenbruch (2001) used a two-part

model; however, the splitting formulation was not consistent with the literature. Further, the model was compared to atypical formulations such as the Wilcoxon, Kolmogorov-Smirnov, and z tests. As such, these types of articles are not included in the subsequent review. One exception is Xie, He, and Goh (2001) who included a likelihood-ratio test for comparing the Poisson and ZIP models.

Statistical

Greene (1994) proposed several 'zero-altered count models' for comparison. First, he took Mullahy's with-zero's (WZ) adaptation of the Hurdle model and included a scalar estimate for ease on computational burdens. Greene also presented an adaptation of Lambert's ZIP known as ZIP(τ) and modified it for the negative binomial formulations terming them ZINB and ZINB(τ). The intention was to identify a "procedure which will enable us to test the zero inflated model against the simple Poisson model or against the negative binomial model. The latter will allow us to make a statement as to whether the excess zeros are the consequence of the splitting mechanism or are a symptom of unobserved heterogeneity" (Greene, 1994, p.10). Greene developed a method for comparing the models; however, he noted that there was no a priori reason to think that the Vuong statistic would be inferior.

Applications

Zorn (1996) examined the counts of actions taken by Congress addressing Supreme Court decisions between 1953 and 1987. The zeros were seen to arise from two sources since many cases will not be addressed unless there are lobbyists to pressure redress. Covariates included the year of the decision, the political orientation of the decision, the presence of lower court disagreement, the presence of precedence alteration, declaration of unconstitutionality, and unanimous vote. The number of actions ranged from 0 to 11 (M = 0.11, SD = .64); however, 3,882 (95.8%) of the 4,052 counts were zeros. This contributed to an exceptionally high skew of

7.97. When ignoring the zeros, the skew was reduced to 1.86 (M = 2.59, SD = 1.53). The observed zeros were 107% of that which would be Poisson-expected.

Regardless, Poisson model results were in line with theory-driven expectations. However, the test of overdispersion was significant when comparing the Poisson and negative binomial Poisson resulting in fewer significant predictors than if ignoring overdispersion. The author also fitted a generalized negative binomial model in which "the variance parameter is allowed to vary as an exponential function of the same independent variables included in the model of the count" (Zorn, 1996, p.9), which led to even better model fit. However, due to zero-inflation, no model provided reasonable estimate sizes given the low mean count.

Their analyses using the ZIP and Hurdle models yielded several findings. First, the probability of remaining a zero in the transition stage was considerably lower for the Hurdle model than for the ZIP model at lower levels of a predictor. This is a reflection of the asymmetry of the Hurdle model. Second, parameter estimates and standard errors were similar between the two models. They concluded that "at least in some circumstances the performance of ZIP and hurdle Poisson models will be quite similar. This suggests that, as a practical matter and barring any strong theoretical considerations favoring one over the other, the choice between them may be made largely on the basis of convenience of estimation" (Zorn, 1996, p.11).

Pardoe and Durham (2003) compared the Poisson, Hurdle, and ZIP models as well as their negative binomial formulations using wine sales data. Of the 1,425 counts, 1,000 (70.2%) were zeros. The authors noted that this is greater than the 67.8% that a Poisson distribution is capable of predicting. Based on the AIC, the zero-inflated negative binomial ZIP performed best. Surprisingly, the Hurdle model fit more poorly than did the Poisson model. It is possible, given the unrealistically high AIC relative to the other models, that the value wasn't calculated

correctly. Alternatively, the distribution may not have been correctly specified since their analysis included Bayesian estimates of the prior distributions. No discussion pertaining to the Hurdle model was included. However, they did provide a novel procedure for comparing the range of fit statistics across the zero-inflation models. This 'parallel coordinate plot for goodness of fit measures' consists of an x-axis labeled Min on the left and Max on the right. The y-axis is a series of horizontal lines each pertaining to a fit statistic (e.g., AIC, BIC). The ceiling x-axis contains the labels for the models being compared. Then, points for each model are plotted on the lines for the fit statistics at their relative location between Min and Max. A similar procedure restricted to the AIC was used by Warton (2005). This technique was adapted to display the coverage for simulated fit statistics.

Welsh, Cunningham, Donnelly, and Lindenmayer (1996) used zero-inflated rare species count data to compare the Poisson, negative binomial Poisson, Hurdle, negative binomial Hurdle, and ZIP models. Approximately 66% of the observations were zeros. They found little difference between the Hurdle, negative binomial Hurdle, and ZIP model results. Since there was no overdispersion, the authors recommended using the Poisson model. This is in line with Warton's (2005) assertion that the more complex zero-inflation models may not always be necessary; at least, this appears to be the case with 66% zero-inflation and equidispersion.

Discrepant Findings

What is exactly meant by zero-inflation? Min and Agresti (2005) define zero-inflated count data as "data for which a generalized linear model has lack of fit due to disproportionately many zero" (p.1). This raises the question, "At what point does the frequency of zeros become disproportionate to the frequency of non-zeros?" One statistical definition states that the proportion of zeros is greater than that to be expected given the posited distribution (Zorn, 1996). For example, for count data, the proportion of zeros should not be greater than that expected by a

Poisson distribution. However, there are three problems with this. First, there may be many different proportions of zeros greater than that expected by a particular Poisson distribution. Second, the definition assumes that the full model is distributed Poisson. Third, it ignores the two potential sources of overdispersion for Poisson zero-inflated data.

The aforementioned AEWR example displayed a zero proportion of .7158 with a mean of .54, a standard deviation of 1, and a skew of 1.971. Ignoring the zeros, although this event stage distribution remains negative skewed, the mean increased to 1.91, and the level of skew dropped to 0.96. The distribution for the categorical sex variable was binomial with approximately 43% males and 47% females. The distribution for the age variable was roughly normal with a mean of 48.86, a standard deviation of 17.41, and a skew 0.87. Hence, with 71% zeros, a heavily skewed distribution of 1.971, a moderately skewed nonzero distribution of 0.96, a normally distributed continuous predictor, and a two-level categorical predictor led to the following findings: 1) the Hurdle model fit better than Poisson model; 2) the negative binomial Hurdle fit better than negative binomial Poisson model; 3) the negative binomial Hurdle fit better than the Hurdle model; 4) the negative binomial ZIP fit better than ZIP model; 5) the negative binomial ZIP model descriptively fit better than all others; and, 6) the Hurdle and negative binomial Hurdle model yielded nearly identical estimates and *p*-values. Hence, findings between the zero-inflation models differed in terms of both fit and the significance of parameter estimates.

Although not all research presented sufficient information (e.g., data necessary to calculate skew), there is clearly enough variation in results to warrant further research. Mullahy's (1986) Hurdle model analyses were impacted by zero-inflation of .263 and not by zero-inflation of .616 or .404; however, this is in disagreement with findings that the Hurdle model adequately handles zero-deflation (Min, 2003). Lambert's ZIP analysis with 71.8% zeros favored the ZIP over the

negative binomial ZIP. Greene's (1994) ZIP analyses resulted in nonsensical results under .894 zeros and heavy skew (4.02); the negative binomial ZIP corrected this. Slymen, Ayala, Arredondo, and Elder's (2006) ZIP and negative binomial ZIP results were virtually identical; however, their event stage distribution was uniform. This was confirmed by Warton's (2005) finding that the negative binomial fits better than the ZIP only when zero-inflation and overdispersion both are indeed present. Extending from this is Böhning, Dietz, Schlattmann, Mendonça, and Kirchner's (1999) findings that the ZIP actually fit better than the Poisson given .216 and .289 zero-deflation. This is again in contrast to the suggestion that the Hurdle, and not the ZIP, is appropriate for zero-deflated data. However, it could be argued that a normal distribution should have been assumed given that the event stage distribution was relatively normal.

When comparing the Hurdle model to the ZIP model, Zorn (1996) found similar results given .958 zero-inflation, skew of 7.97, and a reduction of skew to 1.86 for the event stage. These findings are in contrast to Zorn's (1996) findings of greater zero-inflation and, subsequently, greater skew. Welsh, Cunningham, Donnelly, and Lindenmayer (1997) also found little difference between the Hurdle and ZIP models. Table 2-2 summarizes the findings from the zero-inflation literature.

There are three factors that may have caused the anomalous findings. The first possibility pertains to the effect of different types, quantities, and values for predictors. The second possibility is the proportion of zeros for the outcome variable. The third possibility is the degree of skew in the event stage for the outcome variable.

It has already been suggested that the Hurdle and ZIP models should be chosen given a priori research about the source and nature of the zeros. Further, it has been established that the

negative binomial formulations are meant to handle additional overdispersion in the event stage. However, the previous findings suggest that there are additional considerations such as the proportion of zeros and the nature of the event stage distribution. The proportion of zeros in this research ranged from as low as .20 (Delucchi & Bostrom, 1994) to .958 (Zorn, 1996). Distributions for the event stage included those that were heavily positively skewed (Greene, 1994), moderately positively skewed (AEWR example), distributed normally (Böhning, Dietz, Schlattmann, Mendonça, & Kirchner, 1999), and distributed uniformly (Slymen, Ayala, Arredondo, & Elder, 2006).

The first possibility, pertaining to the covariates, can only be tested by varying an incredibly large set of conditions ranging from small to large quantities of predictors as well as their types (e.g., nominal, ordinal), and distributions. However, given a particular set of covariates and corresponding values, the other two possibilities pertaining to zero-inflation and skew can be explored. It is possible to vary the proportion of zeros for the outcome variable and to simultaneously vary the degree of skew of the nonzeros for this outcome variable. Hence, the following research questions are presented:

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated log-likelihood between a) the Negative binomial Poisson model vs. Poisson model; b) the Hurdle model vs. Poisson model?; c) the Negative binomial Hurdle model vs. negative binomial Poisson model?; d) the Negative binomial Hurdle model vs. Hurdle model; and, e) the Negative binomial ZIP model vs. ZIP model?

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated AIC between all models?

These questions were answered by establishing several levels of skew and zero-inflation. The covariates and their values were fixed as one continuous variable from a standard normal distribution and one binary variable. Data for the outcome variable, for all levels of skew and

zero-inflation, were simulated to estimate log-likelihood values, AIC indices, covariate coefficients, standard errors, and *p*-values. The details are delineated in the following chapter.

The objective is consistent with Zorn's (1996) advice: "First and foremost, work should be undertaken to better ascertain the statistical properties of the various estimators outlined here. It is important that we determine the robustness of these techniques to skewness in the dependent variable, model misspecification, and the host of other problems that all too frequently plague political science researchers . . . perhaps using Monte Carlo methods to assess under what circumstances the results of the two may diverge" (Zorn, 1996, p.12).

Table 2-1. Five pairs of nested models valid for statistical comparison

	Valid Comparisons (Nested Models)				
	1	2	3	4	5
Poisson	Simple			Simple	
NB Poisson	Complex				Simple
Hurdle		Simple		Complex	
NB Hurdle		Complex			Complex
ZIP			Simple		
NB ZIP			Complex		

Table 2-2. Summary of literature on zero-inflation

Researcher(s)	Models Compared	Zeros	Superior Model	Comments
Min & Agresti (2004)	Hurdle vs. ZIP	Simulation	Hurdle	Zero-deflation
Min & Agresti (2004)	Hurdle vs. ZIP	Simulation	Equal	Zero-inflation
Mullahy (1986)	Hurdle	.26, .62, .41	Hurdle .26	
Lambert (1992)	ZIP vs. NB Poisson vs. Poisson	.718	ZIP over NB Poisson over Poisson	
Greene (1994)	ZIP vs. NB ZIP vs. NB Poisson	.894	NB ZIP over ZIP; NB Poisson over ZIP over Poisson	Heavy skew; Probit link
Slymen, Ayala, Arredondo, and Elder (2006)	Poisson vs. NB Poisson; ZIP vs. NB ZIP	.824	NB Poisson; Equal	Uniform event stage; Overall, AIC's favor ZIP
Xie, He, and Goh (2001)	ZIP vs. Poisson	.87	ZIP	Outliers
Böhning, Dietz, Schlattmann, Mendonça, and Kirchner's (1999)	Poisson vs. ZIP	.216, .289	ZIP	Zero-deflation; normal event stage
Zorn (1996)	Hurdle vs. ZIP	.958	Equal	Heavy skew
Pardoe and Durham (2003)	Poisson vs. Hurdle vs. ZIP	.702	NB ZIP over Poisson over Hurdle	Based on AIC's
Warton (2005)	ZIP vs. NB ZIP	Various	ZIP or NB ZIP	Overdispersion favored NB ZIP
Warton (2005)	ZIP vs. NB ZIP	Very low	ZIP	Rare convergence for NB ZIP
Warton (2005)	NB ZIP vs. OLS vs. NB ZIP vs. Poisson	Various	NB ZIP over OLS/NB ZIP over Poisson	Based on AIC's
Welsh, Cunningham, Donnelly, and Lindenmayer (1997)	Hurdle vs. ZIP	.66	Equal	No overdispersion

CHAPTER 3
METHODOLOGY

The previous chapter fulfilled three objectives. First, it described the statistical models and methods for analyzing count data including that which is zero-inflated. Second, it presented research, both technical and applied, pertaining to three models (Poisson, Hurdle, and ZIP) as well as their negative binomial formulations. Third, it was concluded that there is considerable divergence in findings between models and that such differences should be explored by examining different levels of zero-inflation and skew for the count outcome variable. This led to the following research questions:

Research Questions

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated log-likelihood between a) the Negative binomial Poisson model vs. Poisson model; b) the Hurdle model vs. Poisson model?; c) the Negative binomial Hurdle model vs. negative binomial Poisson model?; d) the Negative binomial Hurdle model vs. Hurdle model; and, e) the Negative binomial ZIP model vs. ZIP model?

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated AIC between all models?

As recommended by Zorn (1996), these questions were answered using a Monte Carlo study in which the proportion of zeros and the skew for the distribution of the event stage counts varied between simulations.

Monte Carlo Study Design

Monte Carlo studies begin in the same manner as other research methods. First, a problem is identified as a research question. Second, the problem is made concrete in the form of a hypothesis or set of hypotheses. Third, the hypotheses are tested using rigorous methods. Fourth, conclusions are drawn from these results. Fifth, the implications and limitations are elucidated for future researchers and practitioners.

For this study, the problem was zero-inflation for count outcomes. The problem was then clarified in the form of the research questions stated above. It is at this point that Monte Carlo studies differ from most other methods. For the Monte Carlo study, no real data is gathered. Rather, a set of samples are generated based on given parameter specifications resulting in a sampling distribution that is considered to be equivalent to that which would have been obtained had this many real participants been available. This is a clear advantage over the true experiment where statistics are used to infer from one sample to an entire population (Mooney, 1997). The Monte Carlo study is not limited to a finite number of participants and is subsequently not prone to violations of asymptotic theory (Paxton, Curran, Bollen, Kirby, & Chen, 2001). Paxton, Curran, Bollen, Kirby, and Chen (2001, p.287) provide a succinct explanation as follows:

> The researcher begins by creating a model with known population parameters (i.e., the values are set by the researcher). The analyst then draws repeated samples of size *N* from that population and, for each sample, estimates the parameters of interest. Next, a sampling distribution is estimated for each population parameter by collecting the parameter estimates from all the samples. The properties of that sampling distribution, such as its mean or variance, come from this estimated sampling distribution.

Similarly, Mooney (1997, p.2) explains,

> Monte Carlo simulation offers an alternative to analytical mathematics for understanding a statistic's sampling distribution and evaluating its behavior in random samples. Monte Carlo simulation does this empirically using *random samples from known populations of simulated data* to track a statistic's behavior. The basic concept is straightforward: If a statistic's sampling distribution is the density function of the values it could take on in a given population, then its estimate is the relative frequency distribution of the values of that statistic that were *actually observed* in many samples drawn from that population.

The Monte Carlo simulations were performed using the R programming language (R Development Core Team, 2006). R is an open-source language based on the commercially

available S-Plus program (Insightful, 2005) and is just one of many programs that can be used to model zero-inflated data.[23]

Monte Carlo Sampling

Pseudo-Population

Mooney (1997) explains that defining the population parameters requires defining the pseudo-population. In the pseudo-population, the values for a categorical factor, X_1, with two levels were constant at either 0 or 1; this is representative of a typical two-level categorical factor such as sex. In the dataset, these values alternated. The Excel 2003 random number generator was used to draw a random selection of 1,000 normally distributed values to represent the continuous variable, $X_2 \sim N(0,1)$. This resulted in a pseudo-population of $N = 1{,}000$ where the values for categorical X_1 and continuous X_2 were known. The values for X_2 ranged from -2.945 to 3.28 with a mean of 0, a median of 0.05, a standard deviation of 0.986, and skew of -0.127. These two sets of covariates and their distributions were chosen as a parsimonious generalization to the basic ANCOVA general linear model that extends from analyses with either quantitative or qualitative predictors.[24] The simulations varied in terms of a) the amount of zero-inflation present in the outcome variable scores; b) the amount of skew present in the event stage outcome variable scores, and c) the generalized linear model.

The outcome variable, *Y*, was established as a deterministic variable in that it varied systematically as a function of the specified distributions. As clarified by Mooney (1997), "Deterministic variables are vectors of numbers that take on a range of values in a prespecified, nonrandom manner" (p.6). The regression coefficients for X_1 and X_2 are random variables that

[23] Others include Stata, LIMDEP, COUNT, MATLAB, and SAS. Preliminary analyses compared results between SAS-code incorporated in research (Min & Agresti, 2004; Min, 2003) to publicly available R-code written by Simon Jackman of the Stanford Political Science Computing Laboratory to verify the comparability of results.

[24] The model is similar to that of the AEWR examples.

take on their realizations as a result of the relationship between deterministic *Y* and the two random *X* covariates.

The Prespecified Zero Proportions

Justification for generating values with prespecified proportions of event counts with frequencies also determined by prespecified proportions of zeros is justified by Mooney (1997). "If we know (or are willing to make assumptions about) the components that make up a statistic, then we can simulate these components, calculate the statistic, and explore the behavior of the resulting estimates" (Mooney, 1997, p.67). In his case, the concern was bias determined by calculating statistics and inspecting graphs as a result of Monte Carlo simulations. For this study, the concern was goodness-of-fit for six models by comparing log-likelihoods and AIC's and inspecting graphs as a result of Monte Carlo simulations over three levels of skew and five levels of zero proportions.

Previous research displayed zero-inflation ranging from .20 (Mullahy, 1986) to .96 (Zorn, 1996). To reflect a range including both zero-deflation and zero-inflation, six pseudo-populations were established differing in the proportion of zeros present in the count outcome variable. The pseudo-populations contained either 0.10, 0.25, 0.50, 0.75, or 0.90 proportions of zeros.[25]

Pre-Specified Skew

To manipulate skew, the event stage distributions for the count outcome variable were varied over three conditions. For each condition, proportions were specified and values were drawn randomly from a multinomial distribution such that the frequencies of the values added to the frequency for those already drawn to represent zero-inflation summed to 1,000. In other

[25] Originally, a 0.00 proportion of zeros was included as a control condition. However, for the case of negative skew, this is simply a count model truncated at one. And, for a normal distribution event-stage, this is semi-continuous data often tested with different methods from those for zero-inflation (Min, 2002).

words, if .50 (or 500) observations were assigned a value of zero then the remaining .50 (or 500) observations had values drawn from the prespecified multinomial distribution.

Event stage values ranging from one to five were sampled in order to represent a range small enough to distinguish the distribution from one that might be analyzed as continuous given a particular shape. The prespecified probabilities for each of the five counts were determined primarily to achieve a particular level of skew and secondarily in order to achieve a convergent and admissible solution. Hence, the proportions were not always exactly equal to .10, .25, .50, .75, and .90; the approximates were selected to achieve convergence leading to more trust in convergence for the Monte Carlo simulations. Table 3-1 displays the proportions of each count as a function of the three levels of skew and five levels of zeros. Table 3-2 displays this same information in terms of frequencies instead of proportions. Table 3-3 displays the descriptive statistics for each distribution.

Random Number Generation

By definition, a random number is one in which there is no way possible to a priori determine its value. Most statistical analysis software packages include random number generators. However, these generated random numbers are not truly random. Usually, one specifies a seed; when replications are performed using the same seed, the generated numbers are identical to the first. Hence, the values are pseudo-random (Bonate, 2001). However, this limitation is actually an advantage in that the researcher can check for errors in model programming and run the analysis again with the same generated sample (Mooney, 1997).[26]

[26] Technically, this is only true for most random number generators. The R programming language, by default, bases its generation on the Wichman-Hill algorithm and the system clock resulting in a 626-integer seed (R Development Core Team, 2006).

Another feature of Monte Carlo random sampling pertains to the desired distributions. Typically, the random numbers are drawn from a uniform distribution, which is then followed by a transformation to the desired distribution. As Mooney (1997) explains, "In its standard form, U(0, 1), the uniform distribution is the building block of all Monte Carlo simulation work in that from it, in one way or another, variables with all other distribution functions are derived. This is because the U(0, 1) distribution with its $0 \leq x \leq 1$ range, can be used to simulate a set of random probabilities, which are used to generate other distribution functions through the inverse transformation and acceptance-rejection methods" (p.10).

The random number generation was performed using R 2.3.1 (R Development Core Team, 2006). The procedure requires the generic *sample* command in which the following were specified: 1) a range of counts, which in this study, was from one to five (not zero to five since proportions of zeros were already drawn from the pseudo-population), 2) the number of values to draw, which in this study was one minus the prespecified proportion of zeros, 3) the proportions for each value in the range, which in this case was one of three possibilities determining skew, and 4) the specification to sample with replacement. The seed was arbitrarily set at 6770.

Sample Size

Determining the appropriate sample size for each simulate is an important concern. This could range from zero to infinity. However, if the sample size is too small then it is not safe to assume that estimates are asymptotically normal. On the other hand, computer time and burden increases as sample size increases. The sample size was based on the highest found in the literature pertaining to zero-inflated count data, which was $n = 1{,}000$ (Civentti & Hines, 2005).

Simulation Size

Determining the number of simulations is also an important concern since too few replications may result in inaccurate estimates and too many replications may unnecessarily

overburden computer time and performance (Bonate, 2001). Hur's (1999) research pertaining to the ZIP model with random effects set the number of simulations at 200. Min and Agresti (2005) were able to sufficiently compare the goodness of fit for several competing models using 1,000 simulations. Likewise, Civettini and Hines (2005) selected 1,000 simulations when researching misspecification in negative binomial ZIP models. Lambert (1989) set the number of simulations at 2,000 when researching the asymptotic properties of the ZIP model.

Mooney (1997) states that "The best practical advice on how many trials are needed for a given experiment is "lots"! Most simulations published recently report upward from 1,000 trials, and simulations of 10,000 and 25,000 trials are common" (p.58). Given the previously noted problems with convergence for the negative binomial ZIP model, it seems prudent to minimize the number of simulations as much as possible. However, it is also important to simulate under conditions already found to produce asymptotic results. Hence, similar to Lambert's (1989) seminal study and equal to the maximum found in the literature, the number of simulations was set at 2,000 for each condition ($S = 2{,}000$).

Iteration Size

Iteration size is not much of a concern for the basic Poisson and negative binomial Poisson model. However, obtaining valid estimates for the Hurdle model, ZIP model, and their negative binomial counterparts requires selecting an appropriate maximum number of iterations. Too few iterations can lead to incorrect estimates or, even worse, premature declaration of nonconvergence. Too many iterations results in unnecessary computer time and burden. Various procedures for analyzing these models in R have maximum iterations of 500, 5,000, and 50,000. It was important to determine an iteration size that would be equal across analyses and large enough given some models' potential for nonconvergence. These concerns were deemed more

important than excessive computational burden. Hence, the procedure with the largest iteration size was selected for all procedures leading to 50,000 iterations per analysis.

Distribution Generation

The following describes the procedure for generating the distributions for each simulation.

- The total proportion of counts out of 1,000 to be sampled was reduced by the prespecified proportion of zeros. Hence, if the proportion of zeros was .50 then the proportion of event stage counts was 1.00-0.50 = .50. Translated into frequencies, this is 1,000-(0.50 * 1,000) = 500

- The generic R 'sample' procedure was used to sample with replacement from the event stage counts according the specified proportions depending on the skew condition. The seed was set at 6770.

- The values were sampled over $N = 1{,}000$.

- Each sample was simulated $S = 2{,}000$ times.

- The data over all $S = 2{,}000$ at $N = 1{,}000$ were then stored in separate files as they were created. The filenames conformed to the labeling format where the model was replaced by an underscore (e.g., _25Pos was the filename for the $S = 1{,}000$ datasets at $N = 2{,}000$ where the proportion of zeros was .25 and the skew for the event stage was positive).

Monte Carlo Models

As previously discussed, generalized linear models include a random component, a systematic component, and a link function. The X_1 and X_2 constants form the systematic component in the pseudo-population's generalized linear model. The random component specification for the distribution of the outcome mean varied from pseudo-population to pseudo-population. The base level generalized linear model assuming a normally distributed outcome is given by

$$Y_i = \beta_0 + \beta_1(X_{1i}) + \beta_2(X_{2i}) + \varepsilon_i \,. \tag{3-1}$$

Subsequent models extended from this base model to form the six distributions for deterministic Y. The first model, which was the Poisson generalized linear model with a log link, is given by

$$\log(\lambda_i) = \beta_0 + \beta_1(X_{1i}) + \beta_2(X_{2i}) \quad (3\text{-}2)$$

Table 3-4 displays the parameters for this model over all conditions of skew and zero-inflation. For example, the analysis with a .10 proportion of zeros and a positively skewed event stage distribution yielded

$$\log(\lambda_i) = .450 + .004(X_1) - .037(X_2) \quad (3\text{-}3)$$

For both predictors, the coefficient near zero transforms to an exponentiation near one. The coefficient for X_1 is lowest (.001) for the negatively distributed data with a .25 proportion of zeros and highest (.007) for the positively distributed data with a .50 proportion of zeros. The coefficient for X_2 is lowest (-.007) for the negatively distributed data with a .10 proportion of zeros and highest (-.165) for the normally distributed data with a .90 proportion of zeros. Hence, for the two-level categorical X_1 variable, changing from zero to one multiplies the mean of the outcome variable by approximately exp(0.00), which equals one. For the simulated data, the test for the coefficient estimates corresponding to these pseudo-population parameter values is approximately H_0: $\beta = 0$.

The second model, which was the negative binomial formulation of the Poisson model, is the same as the Poisson model with the addition of a dispersion parameter. Table 3-5 displays the parameters for this model over all conditions of skew and zero-inflation. Like the Poisson model, the test for the simulated data is H_0: $\beta = 0$. The dispersion parameter is also included in Table 3-5. For the simulated data, the test that this parameter equals zero has equivalent results to the model comparison tests conducted.

The third and fifth models were the general formulations for the Hurdle and ZIP models given by

$$\text{logit}(p_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1(Z_{1i}) + \beta_2(Z_{2i}) \quad (3\text{-}4)$$

$$\log(\lambda_i) = \beta_4 + \beta_5(X_{1i}) + \beta_6(X_{2i}) \quad (3\text{-}5)$$

while the fourth and six models were their negative binomial formulations. Tables 3-6 through 3-13 display the pseudo-population coefficients, standard errors, log-likelihood values, and AIC values over all conditions of proportions of zero and skew. Although there are four models (i.e., Hurdle, negative binomial Hurdle, ZIP, and negative binomial ZIP), there are eight tables since, for each model, there are separate results for the transition (zeros) stage and events (nonzero counts) stage.

Monte Carlo Analysis Procedures

A generic looping procedure was written in the R programming language to retrieve each of the 15 sets of simulated data. Each dataset was analyzed with each of the six models. The Poisson model used the generic *glm* procedure in R, which requires specification of the model, Poisson distribution, and log link. The negative binomial Poisson model used the *glm.nb* procedure in R from the generic *MASS* library. This procedure requires only the specification of the model. The R Core Development Team (2006) describes the procedure as follows: "An alternating iteration process is used. For given 'theta' [dispersion] the GLM is fitted using the same process as used by 'glm()'. For fixed means the 'theta' parameter is estimated using score and information iterations. The two are alternated until convergence of both is achieved" (R Core Development Team, 2006).

The Hurdle model used the *hurdle* procedure in the *pscl* library authored by Simon Jackman, PhD of the Political Science Computing Laboratory at Stanford University. The

procedure requires specification of the count response variable (Y), the Poisson distribution for the event stage, the logit link function for the transition stage, and the models for both the transition stage and event stage, which for this study were X_1 and X_2. The negative binomial Hurdle model also used the hurdle procedure but specified a negative binomial rather than a Poisson distribution for the event stage. This hurdle model procedure maximizes the log-likelihood using either the Broyden-Fletcher-Goldfarb-Shanno (BFGS) or Nelder-Mead methods. The Nelder-Mead (default) was selected for solution optimization.

The ZIP model used the *zicounts* procedure in the *zicounts* library authored by Samuel Mwalili, doctoral student in biostatistics at Katholieke Universeteit Leuven (Netherlands). Similar to the hurdle procedure, *zicounts* requires specification of the count outcome variable and models for both the transition stage and event stage. The distribution is specified as "ZIP". Optimization procedures include the BFGS, Nelder-Mead, and conjoint gradient (CG) methods; for consistency, the Nelder-Mead was chosen. The negative binomial ZIP used the same procedure with the specification of the "ZINB" distribution. The specific procedure for analyzing the data with the models was as follows:

- Three separate loops were established for the negatively skewed, normal, and positively skewed distributions.
- Arrays were created to store the log-likelihood, AIC, coefficient estimates, standard errors, and *p*-values.
- The outermost loop for each distribution pertained to the five conditions of varying zero proportions; at each loop, the data corresponding to this zero-condition and distribution was loaded.
- Within these loops, another looping procedure pertained to the six models that analyzed the data.
- Within these loops another looping procedure was defined for the number of simulations in the data set.

- It was at this point that the data corresponding to a particular distribution, proportion of zeros, model, and simulation were analyzed with the calculated AIC, log-likelihood, coefficient estimates, standard errors, and *p*-values transferred to the appropriate array. Models that failed to converge were automatically coded as NA.[27]

- The results for each statistic over all simulations for a particular distribution, model, and proportion of zeros were then exported in comma-separated format for subsequent analyses. Hence, the three distributions by five proportion of zeros conditions by six models yielded 90 data files each containing columns pertaining to a particular statistic or set of statistics and 2,000 rows of simulated results.

Analysis Design

A series of tests were conducted using the simulated data. The results and graphical output were created using R 2.01 (R Development Core Team, 2006) and SPSS 14.0. The design for a particular analysis depended on the research question. These questions were as follows:

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated log-likelihood between a) the Negative binomial Poisson model vs. Poisson model; b) the Hurdle model vs. Poisson model?; c) the Negative binomial Hurdle model vs. negative binomial Poisson model?; d) the Negative binomial Hurdle model vs. Hurdle model; and, e) the Negative binomial ZIP model vs. ZIP model?

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated AIC between all models?

The first question was answered by calculating log-likelihood values for the six models over the six zero proportion conditions and the three skew conditions. The deviance statistic was then calculated as $-2(LL_S-LL_C)$ where LL_S is the model with less parameters (i.e., the simple model) than the LL_C (i.e., the more complex model). Since there were 5 conditions for the proportions of zeros and 3 conditions for skew, this led to a) a total of 15 separate analyses comparing the fit of the negative binomial Poisson model and the Poisson model; b) 15 separate analyses comparing the fit of the Poisson model and the Hurdle model; c) 15 separate analyses comparing the fit of the negative binomial Poisson model and the negative binomial Hurdle

[27] This number was chosen to reduce the probability of obtaining a valid result that would be identified as a nonconvergant solution

model; and d) 15 separate analyses comparing the fit of ZIP model and the negative binomial ZIP model.

Test statistics for the deviance were assumed asymptotically chi-square with degrees of freedom equal to the difference in the number of parameters between the two models. Some models differed only by the dispersion parameter; these were a) the comparison of the Poisson model and negative binomial Poisson model; b) the comparison of the Hurdle model and negative binomial Hurdle model; and c) the comparison of the ZIP model and negative binomial ZIP model. The log-likelihood statistic for the Hurdle model is based on the log-likelihood statistics from each of its two parts. Given three parameters in the Poisson model (i.e., the intercept and the two predictor coefficients), there are six parameters in the Hurdle model. Hence, the degrees of freedom for the model comparison test are the difference, which are three. The same is true for comparing the negative binomial Poisson model to the negative binomial Hurdle model where including the dispersion parameter leads to subtracting 4 parameters from 7 parameters.

Each of the 2,000 goodness-of-fit statistics for the simpler model was subtracted from each of the 2,000 goodness-of-fit statistics for the more complex model. These results were then multiplied by -2. Each of these values was then compared to the chi-square distribution with the appropriate degrees of freedom. This yielded a *p*-value representing the probability of obtaining a statistic this high or higher given that the simpler model adequately fits the data. Values exceeding this critical chi-square statistic based on a Type I error rate of $\alpha = .05$ were coded "1" with these results suggesting that the more complex model fits the data better than the Poisson model. Values failing to exceed the critical value were coded "0" with these results suggesting adequate fit for the simpler model.

The results were thus the proportion of simulated datasets favoring the more complex model over the simpler model for a particular proportion of zeros and level of skew. Output included 1.) descriptive statistics for the goodness-of-fit for each model and 2.) boxplots for the difference in goodness-of-fit between the two models.

Answering the second question was done in a similar fashion to answering the first question. This is due to the fact that the AIC is a linear transformation of the log-likelihood statistic with a result that is positive in sign and is interpreted in a lower-is-better fashion. However, these analyses did not involve comparisons to the chi-square distribution. As previously explained, the AIC should not be used in this manner. However, the advantage is that the AIC can be used to descriptively compare all models regardless of whether one is nested or not within another. Boxplots were created to display the range of AIC's produced over simulations for each analysis.

Table 3-1. Proportions of counts as a function of zeros and skew

Study	Proportion of Zeros	Proportion of Remaining Nonzero Values				
		Ones	Twos	Threes	Fours	Fives
_Pos10	0.10	0.504	0.227	0.088	0.054	0.027
_Norm10		0.099	0.198	0.306	0.198	0.099
_Neg10		0.027	0.054	0.091	0.227	0.501
_Pos25	0.25	0.418	0.190	0.075	0.046	0.021
_Norm25		0.083	0.166	0.254	0.166	0.081
_Neg25		0.022	0.045	0.075	0.188	0.420
_Pos50	.50	0.280	0.125	0.050	0.030	0.015
_Norm50		0.053	0.107	0.175	0.110	0.055
_Neg50		0.015	0.030	0.050	0.125	0.280
_Pos75	0.75	0.140	0.062	0.025	0.015	0.008
_Norm75		0.028	0.052	0.089	0.053	0.028
_Neg75		0.008	0.015	0.025	0.062	0.140
_Pos90	0.90	0.056	0.025	0.010	0.006	0.003
_Norm90		0.011	0.021	0.035	0.024	0.009
_Neg90		0.004	0.005	0.010	0.025	0.056

Table 3-2. Frequencies of counts as a function of zeros and skew

Study	Frequency of Zeros	Frequency of Individual Nonzero Values				
		Ones	Twos	Threes	Fours	Fives
_Pos10	100	504	227	88	54	27
_Norm10		99	198	306	198	99
_Neg10		27	54	91	227	501
_Pos25	250	418	190	75	46	21
_Norm25		83	166	254	166	81
_Neg25		22	45	75	188	420
_Pos50	500	280	125	50	30	15
_Norm50		53	107	175	110	55
_Neg50		15	30	50	125	280
_Pos75	750	140	62	25	15	8
_Norm75		28	52	89	53	28
_Neg75		8	15	25	62	140
_Pos90	900	56	25	10	6	3
_Norm90		11	21	35	24	9
_Neg90		4	5	10	25	56

Table 3-3. Descriptive statistics for each distribution

	Range = 0 to 5			Range = 1 to 5		
	Mean	Std.Dev.	Skew	Mean	Std.Dev.	Skew
_Pos10	1.570	1.127	1.155	1.750	1.051	1.45
_Norm10	2.700	1.414	-0.310	3.000	1.150	0.08
_Neg10	3.820	1.619	-1.356	4.250	1.053	-1.43
_Pos25	1.310	1.181	1.092	1.750	1.046	1.42
_Norm25	2.250	1.635	-0.069	2.990	1.149	0.08
_Neg25	3.190	2.054	-0.673	4.840	0.367	-1.86
_Pos50	0.880	1.149	1.514	1.750	1.053	1.44
_Norm50	1.510	1.710	0.587	3.010	1.140	-0.01
_Neg50	2.130	2.253	0.236	4.250	1.053	-1.44
_Pos75	0.440	0.928	2.577	1.760	1.064	1.44
_Norm75	0.750	1.422	1.660	3.000	1.149	-0.00
_Neg75	1.060	1.914	1.349	4.240	1.064	-1.44
_Pos90	0.180	0.622	4.490	1.750	1.058	1.46
_Norm90	0.300	0.965	3.268	2.990	1.124	-0.06
_Neg90	0.420	1.318	2.908	4.240	1.084	-1.51

Table 3-4. Poisson model: pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	LL	AIC
_Pos10	.450	.036	.004	.050	-.037	.026	-1475.3	2956.6
_Norm10	.992	.027	.003	.038	-.015	.020	-1806.7	3619.3
_Neg10	1.340	.023	.002	.032	-.007	.016	-2019.2	4044.5
_Pos25	0.270	.039	.003	.055	-.024	.028	-1468.1	2942.3
_Norm25	.807	.030	.004	.042	-.016	.021	-1913.9	3833.8
_Neg25	1.159	.025	.001	.035	-.009	.018	-2265.35	4534.7
_Pos50	-.138	.048	.007	.068	-.047	.034	-1334.3	2674.5
_Norm50	.407	.036	.004	.052	-.038	.026	-1888.5	3783.0
_Neg50	.752	.031	.003	.043	-.030	.022	-2370.5	4746.9
_Pos75	-.829	.068	.005	.095	-.077	.048	-972.9	1951.7
_Norm75	-.291	.052	.003	.073	-.079	.037	-1461.8	2929.6
_Neg75	.057	.043	.002	.061	-.044	.031	-1919.7	3845.3
_Pos90	-1.77	.109	.036	.151	-.158	.076	-546.5	1099.0
_Norm90	-1.23	.083	.022	.116	-.165	.058	-852.6	1711.1
_Neg90	-.875	.069	.011	.097	-.149	.049	-1152.1	2310.2

Table 3-5. Negative Binomial Poisson model: pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	LL	AIC
_Pos10	.450	.032	.004	.045	-.037	.023	.45	-1594.5
_Norm10	.991	.023	.003	.033	-.015	.017	.32	-1998.9
_Neg10	1.340	.019	.002	.027	-.007	.014	.24	-2277.9
_Pos25	.269	.040	.004	.057	-.025	.029	.64	-1509.4
_Norm25	.807	.033	.004	.046	-.016	.023	.56	-1931.5
_Neg25	1.159	.029	.001	.041	-.009	.021	.51	-2235.5
_Pos50	-.139	.059	.008	.083	-.048	.042	1.05	-1290.1
_Norm50	.041	.051	.005	.072	-.039	.036	1.11	-1723.0
_Neg50	.751	.048	.004	.067	-.030	.034	1.16	-2041.3
_Pos75	-.829	.095	.005	.134	-.078	.068	1.61	-910.4
_Norm75	-.292	.085	.004	.120	-.078	.061	1.95	-1275.6
_Neg75	.057	.081	.003	.114	-.044	.058	2.26	-1567.4
_Pos90	-1.77	.161	.037	.225	-.159	.113	2.04	-511.6
_Norm90	-1.23	.146	.020	.204	-.165	.103	2.70	-751.1
_Neg90	-.874	.140	.008	.197	-.149	.010	3.38	-957.3

Table 3-6. Hurdle model (zeros): pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	LL	AIC
_Pos10	-2.198	.149	.003	.211	.044	.107	-1409.5	2830.9
_Norm10	-2.198	.149	.001	.210	.044	.107	-1797.1	3606.1
_Neg10	-2.198	.149	.001	.211	.044	.107	-1928.3	3868.7
_Pos25	-1.099	.103	.000	.146	-.019	.074	-1465.5	2943.0
_Norm25	-1.098	.103	-.001	.146	-.020	.074	-1787.6	3587.2
_Neg25	-1.099	.103	.000	.146	-.020	.074	-1897.9	3807.9
_Pos50	.000	.089	.000	.127	.037	.064	-1296.6	2605.1
_Norm50	-.005	.089	.005	.127	.035	.064	-1510.1	3023.2
_Neg50	.000	.089	.000	.127	.036	.064	-1583.9	3179.7
_Pos75	1.099	.103	.002	.146	.046	.074	-865.2	1742.4
_Norm75	1.100	.103	.000	.146	.047	.074	-970.3	1952.6
_Neg75	1.099	.103	.000	.146	.045	.074	-1008.1	2028.3
_Pos90	2.205	.150	-.001	.211	.137	.107	-444.9	901.8
_Norm90	2.206	.150	-.001	.211	.137	.107	-486.4	984.8
_Neg90	2.204	.150	-.004	.211	.138	.106	-503.0	1018.1

Table 3-7. Hurdle model (events): pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	LL	AIC
_Pos10	.212	.051	.008	.072	-.063	.036	-1409.5	2830.9
_Norm10	1.035	.030	.003	.042	-.012	.021	-1797.1	3606.1
_Neg10	1.430	.024	.002	.033	-.003	.017	-1928.3	3868.7
_Pos25	.214	.056	.008	.078	-.062	.041	-1465.5	2943
_Norm25	1.033	.033	.005	.047	-.027	.024	-1787.6	3587.2
_Neg25	1.431	.026	.001	.037	-.016	.019	-1897.9	3807.9
_Pos50	.208	.068	.018	.096	-.060	.049	-1296.6	2605.1
_Norm50	1.035	.040	.011	.057	-.023	.029	-1510.1	3023.2
_Neg50	1.430	.032	.004	.045	-.012	.023	-1583.9	3179.7
_Pos75	.213	.096	.010	.135	-.087	.067	-865.2	1742.4
_Norm75	1.032	.057	.005	.081	-.054	.040	-970.3	1952.6
_Neg75	1.428	.045	.002	.064	-.011	.032	-1008.2	2028.3
_Pos90	.179	.156	.060	.215	-.070	.107	-444.9	901.8
_Norm90	1.106	.091	.018	.128	-.052	.064	-486.4	984.8
_Neg90	1.420	.071	.009	.101	-.028	.051	-503.0	1018.1

Table 3-8. Negative Binomial Hurdle model (zeros): pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	LL	AIC
_Pos10	-2.197	.149	-.001	.211	.044	.107	-1395.7	2803.4
_Norm10	-2.197	.149	.000	.211	.044	.107	-1800.2	3612.3
_Neg10	-2.205	.150	-.001	.212	.047	.108	-1936.1	3884.1
_Pos25	-1.099	.103	.001	.146	-.019	.074	-1455.3	2922.5
_Norm25	-1.099	.103	.000	.146	-.019	.074	-1790.2	3592.3
_Neg25	-1.099	.103	.000	.146	-.019	.074	-1904.4	3820.8
_Pos50	.000	.089	.000	.127	.036	.064	-1288.9	2589.8
_Norm50	.000	.089	-.001	.127	.037	.064	-1511.9	3035.8
_Neg50	.001	.089	-.001	.127	.036	.064	-1588.1	3188.2
_Pos75	1.099	.103	.000	.146	.045	.074	-861.1	1734.3
_Norm75	1.099	.103	.000	.146	.045	.074	-971.2	1954.4
_Neg75	1.099	.103	.000	.146	.045	.074	-1010.3	2032.5
_Pos90	2.206	.150	-.001	.211	.137	.107	-443.3	898.7
_Norm90	2.244	.152	-.029	.213	.181	.107	-487.3	986.6
_Neg90	2.205	.150	-001	.211	.138	.107	-503.9	1019.8

Table 3-9. Negative Binomial Hurdle model (events): pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	Theta	LL	AIC
_Pos10	-.067	.108	.009	.091	-.071	.046	2.039	-1395.7	2803.4
_Norm10	1.033	.030	.004	.043	-.012	.022	165.718	-1800.2	3612.3
_Neg10	1.429	.024	.002	.034	-.002	.017	164.970	-1936.1	3884.1
_Pos25	-.042	.114	.009	.098	-.067	.051	2.234	-1455.3	2922.5
_Norm25	1.030	.033	.005	.047	-.027	.025	166.348	-1790.2	3592.3
_Neg25	1.431	.026	.001	.037	-.015	.019	166.006	-1904.4	3820.8
_Pos50	-.072	.146	.020	.122	-.065	.062	2.030	-1288.9	2589.8
_Norm50	1.037	.041	.006	.057	-.025	.029	166.574	-1511.9	3035.8
_Neg50	1.429	.032	.004	.046	-.012	.023	166.526	-1588.1	3188.2
_Pos75	-.075	.208	.012	.173	-.097	.088	1.950	-861.1	1734.3
_Norm75	1.031	.058	.005	.081	-.054	.041	166.445	-971.2	1954.4
_Neg75	1.427	.046	.002	.064	-.011	.032	166.377	-1010.3	2032.5
_Pos90	-.109	.332	.078	.273	-.081	.140	2.047	-443.3	898.7
_Norm90	1.051	.091	-.051	.130	-.048	.066	91.979	-487.3	986.6
_Neg90	1.421	.072	.007	.102	-.028	.052	166.606	-503.9	1019.8

Table 3-10. ZIP model (zeros): pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	LL	AIC
_Pos10	-27.010	4434.000	11.490	4435.000	-.363	150.700	-1478.3	2968.6
_Norm10	-3.114	.038	.002	.527	.050	.253	-1800.1	3612.1
_Neg10	-2.365	.175	.002	.246	.049	.125	-1931.3	3874.6
_Pos25	-15.460	3834.000	-3.418	.002	1.213	.001	-1471.1	2954.3
_Norm25	-1.379	.132	.008	.186	-.056	.098	-1790.6	3593.1
_Neg25	-1.162	.109	.001	.154	-.025	.078	-1900.9	3813.9
_Pos50	-.890	.198	.051	.270	-.050	.141	-1299.5	2611.1
_Norm50	-.126	.097	.002	.137	.030	.070	-1513.1	3038.3
_Neg50	-.031	.091	.000	.129	.035	.065	-1586.9	3185.7
_Pos75	.610	.142	.010	.198	-.013	.010	-868.2	1748.5
_Norm75	1.032	.057	.005	.081	-.054	.040	-973.3	1958.5
_Neg75	1.079	.110	-.001	.147	.044	.075	-1011.2	2034.3
_Pos90	1.797	.183	.032	.255	.104	.129	-447.9	907.7
_Norm90	2.133	.152	.002	.214	.128	.108	-489.4	990.8
_Neg90	2.188	.071	.007	.101	-.028	.051	-506.0	1024.1

Table 3-11. ZIP Model (events): pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	LL	AIC
_Pos10	.450	.004	.004	.050	-.037	.003	-1478.3	2968.6
_Norm10	1.035	.003	.004	.043	-.013	.002	-1800.1	3612.1
_Neg10	1.430	.024	.002	.033	-.003	.002	-1931.3	3874.6
_Pos25	.270	.040	.003	.055	-.024	.028	-1471.1	2954.3
_Norm25	1.032	.033	.005	.047	-.028	.025	-1790.6	3593.1
_Neg25	1.431	.026	.001	.037	-.016	.019	-1900.9	3813.9
_Pos50	.205	.069	.024	.096	-.063	.050	-1299.5	2611.1
_Norm50	1.039	.040	.007	.057	-.025	.029	-1513.1	3038.3
_Neg50	1.429	.032	.004	.045	-.012	.023	-1586.9	3185.7
_Pos75	.213	.096	.001	.135	-.085	.067	-868.2	1748.5
_Norm75	1.015	.107	.001	.151	.033	.076	-973.3	1958.5
_Neg75	1.428	.045	.002	.064	-.011	.032	-1011.2	2034.3
_Pos90	.180	.155	.057	.216	-.069	.108	-447.9	907.7
_Norm90	1.017	.091	.017	.128	-.052	.065	-489.4	990.8
_Neg90	1.422	.071	.007	.101	-.028	.051	-506.0	1024.1

Table 3-12. Negative Binomial ZIP model (zeros): pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	LL	AIC
_Pos10	-15.280	142.300	-3.880	830.400	.719	102.100	-1479.3	2972.6
_Norm10	-3.110	.375	.0165	.527	.0510	.253	-1801.1	3616.1
_Neg10	-2.370	.175	-.001	.247	.047	.125	-1932.3	3878.7
_Pos25	-17.090	501.400	2.390	473.600	1.700	133.500	-1471.4	2956.7
_Norm25	-1.380	.132	-.002	.186	-.059	.983	-1791.6	3597.1
_Neg25	-1.162	.109	.000	.037	-.016	.019	-1901.9	3817.9
_Pos50	-4.895	3.373	1.352	2.211	-1.086	1.304	-1292.4	2598.8
_Norm50	-.131	.097	.008	.014	.031	.069	-1514.1	3042.3
_Neg50	-.031	.091	.000	.129	.035	.065	-1587.9	3189.7
_Pos75	.130	.364	.015	.271	-.040	.136	-865.1	1744.3
_Norm75	1.015	.107	.001	.151	.033	.076	-974.3	1962.6
_Neg75	1.080	.104	-.002	.147	.044	.075	-1012.2	2038.3
_Pos90	1.462	.386	.047	.298	.098	.151	-447.3	908.7
_Norm90	2.133	.152	.002	.214	.128	.108	-490.4	994.8
_Neg90	2.187	.150	-.001	.212	.136	.107	-507.0	1028.1

Table 3-13. Negative Binomial ZIP model (events): pseudo-population parameters

	β_0	s_{β_0}	β_1	s_{β_1}	β_2	s_{β_2}	Theta	LL	AIC
_Pos10	.450	.036	.004	.050	-.037	.026	14.69	-1479.3	2972.6
_Norm10	1.040	.030	.004	.043	-.013	.021	12.64	-1801.1	3616.1
_Neg10	1.430	.024	.001	.033	-.003	.017	13.02	-1932.3	3878.7
_Pos25	.270	.040	.003	.057	-.024	.029	3.16	-1471.4	2956.7
_Norm25	1.030	.033	.005	.047	-.027	.025	10.88	-1791.6	3597.1
_Neg25	1.431	.026	.001	.037	-.016	.019	15.94	-1901.9	3817.9
_Pos50	-.127	.067	.045	.108	-.080	.052	.57	-1292.4	2598.8
_Norm50	1.039	.040	.007	.057	-.026	.029	11.70	-1514.1	3042.3
_Neg50	1.429	.032	.004	.045	-.012	.023	14.19	-1587.9	3189.7
_Pos75	-.066	.204	.010	.172	-.099	.087	.69	-865.1	1744.3
_Norm75	1.033	.057	.005	.081	-.054	.040	12.02	-974.3	1962.6
_Neg75	1.430	.045	.000	.064	-.010	.032	12.99	-1012.2	2038.3
_Pos90	-.102	.327	.072	.273	-.080	.140	.73.	-447.3	908.7
_Norm90	1.016	.091	.017	.128	-.052	.065	12.76	-490.4	994.8
_Neg90	1.422	.071	.006	.101	-.028	.051	14.73	-507.0	1028.1

CHAPTER 4
RESULTS

This chapter presents the results based on the previously discussed methods and procedures for analyzing data with varying proportions of zeros and varying event stage distributions. First, the results are presented using the data for the pseudo-population in which the proportions for each count level are exactly that which was rando mly sampled from in the simulations. Tables and figures are included to support interpretation. Second, the results are presented outlined by the skew level (i.e., positive, normal, and negative) and by the proportion of zeros within that skew level (i.e., .10, .25, .50, .75, and .90). For each combination of conditions, the results are presented for the five model comparisons. Third, the results are summarized separately for the negative, normal, and positive event count distributions. Tables and figures are included to support interpretation. The primary purpose of the results was to assist in answering the research questions, which were as follows:

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated log-likelihood between a) the Negative binomial Poisson model vs. Poisson model; b) the Hurdle model vs. Poisson model?; c) the Negative binomial Hurdle model vs. negative binomial Poisson model?; d) the Negative binomial Hurdle model vs. Hurdle model; and, e) the Negative binomial ZIP model vs. ZIP model?

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated AIC between all models?

Pseudo-Population Results

Before addressing the results of the Monte Carlo simulations, it is necessary to discuss the results when each model was analyzed with the pseudo-population data. The prespecified proportions for each count were displayed in Table 3-1. Table 3-2 displayed this same information in terms of frequencies instead of proportions. Table 3-3 displayed the descriptive statistics for each distribution. Tables 3-4 through 3-13 displayed the coefficients, standard

errors, log-likelihood values, and AIC values with each table pertaining to either a specific model or one of the two stages for a specific model.

The following sections provide the results when comparing models using the pseudo-population data. First, the Poisson model results were compared to the negative binomial Poisson model results, both descriptively via AIC's and inferentially via the deviance model comparison test. Second, the results are presented in a likewise manner for the three Hurdle model comparisons. These comparisons were a) the Hurdle model vs. the negative binomial Hurdle model, b) the Poisson model vs. the Hurdle model, and c) the negative binomial Poisson model vs. the negative binomial Hurdle model. Third, the results are presented for the comparison of the ZIP model and the negative binomial ZIP model. Fourth, descriptive results are presented comparing all models for each of the five proportions of zeros.

Pseudo-Population Poisson Models

Based on the AIC, for all proportions of zeros the data fit the Poisson model better when the distribution was positively skewed than when normally distributed. In addition, this data fit the Poisson model better when the distribution was normally distributed than when it was negatively skewed. For example, for the negatively skewed distribution with a .10 proportion of zeros, the AIC was 4,044.5. As the curve shifted left to a normal distribution, the AIC dropped to 3,619.3, and as the curve shifted further left to a positively skewed distribution, the AIC dropped to 2,956.3.

The same pattern emerged for the negative binomial Poisson models. For example, for the negatively skewed distribution with a .10 proportion of zeros, the AIC was 4,563.8. As the curve shifted left to a normal distribution, the AIC dropped to 4,005.9, and as the curve shifted further left to a positively skewed distribution, the AIC dropped to 3,196.9.

For the .10 proportion of zeros condition, the Poisson models had a lower AIC than that calculated for the negative binomial Poisson model. For the .25 proportion of zeros condition, the AIC's were approximately equal between the two models. For the .50, .75, and .90 proportions of zeros conditions, the AIC was lower for the negative binomial Poisson model than for the Poisson model.

The deviance statistic was calculated comparing the Poisson log-likelihood and negative binomial Poisson log-likelihood for each skew condition and zero proportion condition. These statistics are displayed in Table 4-1. For all analyses, the Poisson model is nested within its negative binomial Poisson formulation differing by one degree of freedom (i.e., the dispersion parameter in the negative binomial Poisson model). Hence, at the .05 Type I error rate and assuming deviance statistics asymptotically distributed chi-square, a deviance exceeding 3.84 suggests better fit for the more complex negative binomial Poisson model. This was the result for all analyses in which the proportion of zeros was .50 or greater. However, for all analyses with .10 and .25 proportions of zeros, the statistic was significant in favor of the Poisson model.

Pseudo-Population Hurdle Models

Hurdle vs. Negative Binomial Hurdle

For the Hurdle models, regardless of the proportion of zeros, the positively skewed distributions had a lower AIC than did the normal distributions. These, in turn, had a lower AIC than the negatively skewed distribution. The same was true for the negative binomial Hurdle models. However, for both models, the difference between AIC's for the three skew conditions decreased as the proportion of zeros decreased (i.e., the models became more similar in fit). When comparing the Hurdle models and negative binomial Hurdle models, the AIC's appear to be similar regardless of the proportion of zeros and regardless of skew.

The deviance statistic was calculated comparing the Hurdle log-likelihood and negative binomial Hurdle log-likelihood for each skew condition and zero proportion condition. These statistics are displayed in Table 4-2. For all analyses, the Hurdle model is nested within its negative binomial Hurdle formulation differing by one degree of freedom (i.e., the dispersion parameter in the negative binomial Hurdle model). Hence, at the .05 Type I error rate and assuming deviance statistics asymptotically distributed chi-square, a deviance exceeding 3.84 suggests better fit for the more complex negative binomial Hurdle model.

For the positively skewed distributions, the negative binomial Hurdle model fit significantly better than the Hurdle model, except when the proportion of zeros was .90. In this case the deviance statistic did not exceed 3.84. For the normal distributions, the Hurdle model fit significantly better than the negative binomial Hurdle model when the proportion of zeros was .10 or .25. When the proportion of zeros was .50, .75, or .90, the deviance did not exceed 3.84. For the negatively skewed distributions, the Hurdle model fit significantly better than the negative binomial Hurdle model when the proportion of zeros was .10, .25, .50, or .75. When the proportion of zeros of .90, the deviance statistic did not exceed 3.84.

Poisson vs. Hurdle

Deviance statistics were also calculated to compare the Poisson log-likelihood and Hurdle log-likelihood for each skew condition and zero proportion condition. These statistics are displayed in Table 4-3. For all analyses, the Poisson model is nested within the Hurdle model differing by three degrees of freedom. Hence, at the .05 Type I error rate and assuming deviance statistics asymptotically distributed chi-square, a deviance exceeding 7.82 suggests better fit for the more complex Hurdle model.

The deviances were large supporting the Hurdle model fit over the Poisson model fit. In fact, several of the deviances were over 1,000. There was only one analysis that did not favor the

Hurdle model fit over the Poisson model fit. This was the deviance calculated for the positively skewed distribution with a .25 proportion of zeros.

Negative Binomial Poisson vs. Negative Binomial Hurdle

Deviance statistics were also calculated to compare the negative binomial Poisson log-likelihood and the negative binomial Hurdle log-likelihood for each skew condition and zero proportion condition. These statistics are displayed in Table 4-4. For all analyses, the negative binomial Poisson model is nested within the negative binomial Hurdle model differing by three degrees of freedom (i.e., the duplication of parameters in the negative binomial Hurdle model to represent both a transition stage and an event stage).[28] Hence, at the .05 Type I error rate and assuming deviance statistics asymptotically distributed chi-square, a deviance exceeding 7.82 suggests better fit for the more complex negative binomial Hurdle model.

As was the case when comparing the Poisson model and the Hurdle model, the deviances were large. Likewise, there was one deviance that did not exceed the critical value. However, it was for the positively skewed distribution with .50 zeros rather than the positively skewed distribution with .25 zeros.

Pseudo-Population ZIP Models

The pseudo-population results for the ZIP models were rather similar to those obtained for the Hurdle models. Graphically, regardless of skew condition and proportions of zeros, the AIC's for the ZIP models and negative binomial ZIP models were very similar. Additionally, the AIC's appeared to be equal between the .10 and .25 proportions of zeros conditions.

The deviance statistic was calculated comparing the ZIP log-likelihood and negative binomial ZIP log-likelihood for each skew condition and zero proportion condition. These

[28] The presence of a dispersion parameter is now redundant between models.

statistics are displayed in Table 4-5. For all analyses, the ZIP model is nested within its negative binomial ZIP formulation differing by one degree of freedom (i.e., the dispersion parameter in the negative binomial ZIP model). Hence, at the .05 Type I error rate and assuming deviance statistics asymptotically distributed chi-square, a deviance exceeding 3.84 suggests better fit for the more complex negative binomial ZIP model.

The deviance was significant for only two comparisons. When the distribution was positively skewed, the fit was significantly better for the negative binomial ZIP model than for the ZIP model when the proportion of zeros was either .50 or .75. All other results suggested that the ZIP model fit was adequate.

Comparing AIC's For All Models

For the .10 proportion of zeros condition with negative skew, the AIC's were approximately equal for all models except for the Poisson models. The AIC for the Poisson model was higher than for the other models, and the AIC for the negative binomial Poisson model was highest of all. When the distribution was normal, the only model to have a noticeably different AIC was the negative binomial Hurdle, which was again highest of all. For the positively skewed distribution, there was some separation between the ZIP and Hurdle models, with the Hurdle models having the lower AIC. The differences between these models and their negative binomial formulations appeared to be trivial. The Poisson model appeared to have the same AIC as those displayed for the ZIP models. The AIC for the negative binomial was again highest of all. Between the three distributions, the AIC declined from the negatively skewed distribution to the normal distribution to the positively skewed distribution.

For the .25 proportion of zeros, there was little distinction between the Poisson models and negative binomial Poisson models for all distributions. Further, there was no distribution displaying a nontrivial distinction between the Hurdle models and the ZIP models. For the

positively skewed distribution, all six distributions appeared approximately equal with a slightly higher AIC apparent for the negative binomial Poisson model. Between the three distributions, the AIC's appeared equal for the negatively skewed distribution and the normal distribution; However, the AIC's for the normal distribution were considerably smaller than the AIC's for the other two distributions.

For the .50 proportion of zeros condition with negative skew, the results for the Poisson model and the negative binomial Poisson model reversed. For the negatively skewed distribution, the AIC for the negative binomial Poisson was higher than those for the ZIP and Hurdle models, while the AIC for the Poisson model was higher yet. As in the .50 proportion of zeros condition, the AIC appeared equal for the ZIP and Hurdle models. Also, as in the .50 proportion of zeros condition, there appeared to be no difference between any of the models for the positively skewed distribution. Between the three distributions, the AIC's declined from the negatively skewed distribution to the normal distribution to the positively skewed distribution with declines being most rapid for the Poisson and negative binomial Poisson models.

Beyond the fact that the overall AIC was lower for the .90 proportion of zeros condition than for the .75 condition, these last two conditions displayed similar results. For each distribution, the Hurdle and ZIP model AIC's were approximately equal. They declined slightly from the negatively skewed distribution to the normal distribution and declined a bit more from this distribution to the positively skewed distribution. For the negatively skewed distribution, the negative binomial Poisson AIC was considerably higher than the AIC's for the ZIP and Hurdle models; the AIC for the Poisson model was highest of all. For the normal distribution, both the Poisson and negative binomial Poisson model AIC's declined by approximately 50% of their value for the negatively skewed distribution. For the positively skewed distribution, they

declined further yet revealing a Poisson model AIC slightly higher than that of the negative binomial AIC model, which was slightly higher than that of the nearly equal Hurdle and ZIP model AIC's.

Monte Carlo Simulation Results

Positively Skewed Distribution

These results pertain to the data in which 100 of the 1,000 observations in each of the 2,000 simulated datasets were constant at zero. The remaining 900 values were sampled from the positively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .10 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -271.5 to -201.98 with a mean of -238.24 and a standard deviation of 10.43. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 model comparisons. The average AIC for the Poisson model was 2957.06, while the average AIC for the negative binomial Poisson model was 3193.3, which descriptively supports the inferential findings of better fit for the Poisson model.

For the .10 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 5.14 to 69.8 with a mean of 27.93 and a standard deviation of 9.0. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 model comparisons. The average AIC for the Hurdle model was 2838.81 while the average AIC for the negative binomial Hurdle model was 2804.88, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .10 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 86.36 to 187.64 with a mean of 132.25 and a standard deviation of 14.58. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000

valid model comparisons. The average AIC for the Poisson model was 2957.06 while the average AIC for the Hurdle model was 2838.81, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .10 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 332.81 to 470.81 with a mean of 398.42 and a standard deviation of 19.86. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 model comparisons. The average AIC for the negative binomial Poisson model was 3193.3 while the average AIC for the negative binomial Hurdle model was 2804.88, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .10 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -2.02 to -2.00 with a mean of -2.00 and a standard deviation of .001 . Given a Type I error rate of .05, the deviance statistic was significant for none of the 112 valid model comparisons. The average AIC for the ZIP model was 2967.85 the average AIC for the negative binomial ZIP model was 2984.79, which descriptively suggests better fit for the ZIP model.

Table 4-6 summarizes the log-likelihood statistics for the positively skewed distribution model comparisons with a .25 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 2622.53 and a maximum of 3096.30 across all 9,678 valid model comparisons. Between the six models, the minimum was 2804.88 (i.e., the negative binomial Hurdle model), and the maximum was 3193.30 (i.e., the negative binomial Poisson model). The rank order from lowest AIC to highest AIC was as follows: negative binomial Hurdle, Hurdle, Poisson, ZIP, negative binomial ZIP, negative binomial Poisson. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-7 displays the descriptive statistics for these two models. Figure 4-1 is a boxplot illustrating these results.

The following set of results pertains to the data in which 250 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 750 values were sampled from a positively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .25 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -126.64 to -31.15 with a mean of -83.2 and a standard deviation of 12.3. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 model comparisons. The average AIC for the Poisson model was 2939.65 while the average AIC for the negative binomial Poisson model was 3020.85, which descriptively supports the inferential findings of better fit for the Poisson model.

For the .25 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from .87 to 22.71 with a mean of 5.76 and a standard deviation of 3.47. Given a Type I error rate of .05, the deviance statistic was significant for 1,994 of the 2,000 valid model comparisons. The average AIC for the Hurdle model was 2947.9 while the average AIC

for the negative binomial Hurdle model was 2921, which descriptively supports the 99.7% inferential findings of better fit for the negative binomial Hurdle model.

For the .25 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from .07 to 22.7 with a mean of 5.76 and a standard deviation of 3.47. Given a Type I error rate of .05, the deviance statistic was significant for 503 of the 2,000 valid model comparisons. The average AIC for the Poisson model was 2939.65 while the average AIC for the Hurdle model was 2947.9, which descriptively supports better fit for the Poisson model.

For the .25 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 75.94 to 156.11 with a mean of 109.85 and a standard deviation of 12.33. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the negative binomial Poisson model was 3020.9, while the average AIC for the negative binomial Hurdle model was 2921, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .25 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -10.55 to 17.14 with a mean of .54 and a standard deviation of 3.86. Given a Type I error rate of .05, the deviance statistic was significant for 344 of the 1,850 valid model comparisons. The average AIC for the ZIP model was 2951.12 while the average AIC for the negative binomial ZIP model was 2956.15 , which descriptively supports better fit for the ZIP model.

Table 4-8 summarizes the log-likelihood statistics for the positively skewed distribution model comparisons with a .25 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the

simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and n represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 2754.05 and a maximum of 3127.33 across all valid model comparisons. Between the six models, the minimum was 2921.00 (i.e., the negative binomial Hurdle model), and the maximum was 3020.85 (i.e., the negative binomial Poisson model). The rank order from lowest AIC to highest AIC was as follows: negative binomial Hurdle, Poisson, Hurdle, ZIP, negative binomial ZIP, negative binomial Poisson. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-9 displays the descriptive statistics for these two models. Figure 4-2 is a boxplot illustrating these results.

The following set of results pertains to the data in which 500 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 500 values were sampled from a positively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .50 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 31.23 to 142.04 with a mean of 87.9 and a standard deviation of 16.37. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 2672.88 while the average AIC for the negative binomial Poisson model was 2582.98, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .50 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from .84 to 43.88 with a mean of 15.69 and a standard deviation of 6.8. Given a Type I error rate of .05, the deviance statistic was significant for 1,961 of the 2,000 valid

model comparisons. The average AIC for the Hurdle model was 2601.93 while the average AIC for the negative binomial Hurdle model was 2588.24, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .50 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 35.5 to 142.74 with a mean of 76.95 and a standard deviation of 17.08. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 2674.5 while the average AIC for the Hurdle model was 2601.93, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .50 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from .12 to 18.73 with a mean of 4.74 and a standard deviation of 2.87. Given a Type I error rate of .05, the deviance statistic was significant for 268 of the 2,000 valid model comparisons. The average AIC for the negative binomial Poisson model was 2582.98 while the average AIC for the negative binomial Hurdle model was 2588.24, which descriptively supports better fit for the negative binomial Poisson model.

For the .50 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -1.16 to 38.97 with a mean of 13.61 and a standard deviation of 6.51. Given a Type I error rate of .05, the deviance statistic was significant for 1,894 of the 2,000 valid model comparisons. The average AIC for the ZIP model was 2607.84 while the average AIC for the negative binomial ZIP model was 2596.23 , which descriptively supports the 94.7% inferential findings of better fit for the negative binomial ZIP model.

Table 4-10 summarizes the log-likelihood statistics for the positively skewed distribution model comparisons with a .50 proportion of zeros. In this table, μ_{LL} represents the mean log-

likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 2451.80 and a maximum of 2862.94 across all 10,000 valid model comparisons. Between the six models, the minimum was 2582.09 (i.e., the negative binomial Poisson model), and the maximum was 2672.88 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: negative binomial Poisson, negative binomial Hurdle, negative binomial ZIP, Hurdle, ZIP, Poisson. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-11 displays the descriptive statistics for these two models. Figure 4-3 is a boxplot illustrating these results.

The following set of results pertains to the data in which 750 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 250 values were sampled from a positively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .75 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 68.65 to 180.23 with a mean of 125.85 and a standard deviation of 16.52. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 1955.22, while the average AIC for the negative binomial Poisson model was 1827.36, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .75 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -1.45 to 32.0 with a mean of 8.3 and a standard deviation of 4.77.

Given a Type I error rate of .05, the deviance statistic was significant for 1,963 of the 2,000 valid model comparisons. The average AIC for the Hurdle model was 1749.71 while the average AIC for the negative binomial Hurdle model was 1735.4, which descriptively supports the 98.1% inferential findings of better fit for the negative binomial Hurdle model.

For the .75 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 95.23 to 363.75 with a mean of 219.51 and a standard deviation of 35.63. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 1955.22 while the average AIC for the Hurdle model was 1749.7, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .75 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 35.6 to 186.3 with a mean of 101.96 and a standard deviation of 19.64. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the negative binomial Poisson model was 1827.36 while the average AIC for the negative binomial Hurdle model was 1735.4, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .75 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -2.0 to 29.66 with a mean of 6.27 and a standard deviation of 4.73. Given a Type I error rate of .05, the deviance statistic was significant for 1,304 of the 2,000 valid model comparisons. The average AIC for the ZIP model was 1747.72, while the average AIC for the negative binomial ZIP model was 1743.44, which descriptively supports the 65.2% inferential findings of better fit for the negative binomial ZIP model.

Table 4-12 summarizes the log-likelihood statistics for the positively skewed distribution model comparisons with a .75 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 1617.27 and a maximum of 2188.38 across all valid model comparisons. Between the six models, the minimum was 1735.40 (i.e., the negative binomial Hurdle model), and the maximum was 1827.36(i.e., the negative binomial Poisson model). The rank order from lowest AIC to highest AIC was as follows: negative binomial Hurdle, negative binomial ZIP, ZIP, Hurdle, negative binomial Poisson, Poisson. These rankings are not in line with the rankings for the pseudo-population AIC's; in the pseudo-population, the Hurdle model has a lower AIC than the negative binomial ZIP model. Table 4-13 displays the descriptive statistics for these two models. Figure 4-4 is a boxplot illustrating these results.

The following set of results pertains to the data in which 900 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 100 values were sampled from a positively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .90 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 29.24 to 188.86 with a mean of 69.82 and a standard deviation of 12.5. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 1098.79 while the

average AIC for the negative binomial Poisson model was 1026.98, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .90 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -4.1 to 17.15 with a mean of 2.96 and a standard deviation of 2.69. Given a Type I error rate of .05, the deviance statistic was significant for 611 of the 1,992 valid model comparisons. The average AIC for the Hurdle model was 906.36 while the average AIC for the negative binomial Hurdle model was 897.5, which descriptively supports the 30.7% inferential findings of better fit for the negative binomial Hurdle model.

For the .90 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 86.47 to 367.79 with a mean of 206.43 and a standard deviation of 41.18. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 1098.79 while the average AIC for the Hurdle model was 906.36, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .90 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from 55.21 to 251.73 with a mean of 139.66 and a standard deviation of 28.92. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,992 valid model comparisons. The average AIC for the negative binomial Poisson model was 1026.98 while the average AIC for the negative binomial Hurdle model was 897.5, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .90 proportion of zeros condition with a positively skewed distribution, the deviance statistic ranged from -2.11 to 14.85 with a mean of .96 and a standard deviation of 2.66.

Given a Type I error rate of .05, the deviance statistic was significant for 265 of the 1,992 valid model comparisons. The average AIC for the ZIP model was 904.34 while the average AIC for the negative binomial ZIP model was 905.38, which descriptively supports better fit for the ZIP model.

Table 4-14 summarizes the log-likelihood statistics for the positively skewed distribution model comparisons with a .90 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 834.28 and a maximum of 1313.41 across all valid model comparisons. Between the six models, the minimum was 897.50 (i.e., the negative binomial Hurdle model), and the maximum was 1098.79 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: negative binomial Hurdle, ZIP, negative binomial ZIP, Hurdle, negative binomial Poisson, Poisson. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-15 displays the descriptive statistics for these two models. Figure 4-5 is a boxplot illustrating these results.

Normal Distribution

The following set of results pertains to the data in which 100 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 900 values were sampled from a normal pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .10 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -425.53 to -347.98 with a mean of -385.1 and a standard deviation of 11.5. Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000 valid model comparisons. The average AIC for the Poison model was 3618.76 while the average AIC for the negative binomial Poisson model was 4005.86, which descriptively supports the inferential findings of better fit for the Poisson model.

For the .10 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -40.51 to -4.94 with a mean of -6.61 and a standard deviation of 2.75. Given a Type I error rate of .05, the deviance statistic was significant for none of the 1,503 valid model comparisons. The average AIC for the Hurdle model was 3605.03 while the average AIC for the negative binomial Hurdle model was 3613.55, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .10 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 9.87 to 31.88 with a mean of 19.73 and a standard deviation of 2.94. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 3618.76 while the average AIC for the Hurdle model was 3605.03, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .10 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 359.25 to 439.66 with a mean of 397.78 and a standard deviation of 12.5. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,503 valid model comparisons. The average AIC for the negative binomial Poisson model was 4005.86 while the

average AIC for the negative binomial Hurdle model was 3613.55, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .10 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -16.62 to -1.83 with a mean of -2.04 and a standard deviation of .7 . Given a Type I error rate of .05, the deviance statistic was significant (ZIP) for 7 of the 1,982 valid model comparisons. The average AIC for the ZIP model was 3611.05 while the average AIC for the negative binomial ZIP model was 3615 , which descriptively supports the 99.6% inferential findings of better fit for the ZIP model.

Table 4-16 summarizes the log-likelihood statistics for the normal distribution model comparisons with a .10 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples

A comparison of AIC's over all models revealed a minimum of 3530.45 and a maximum of 4071.95 across all valid model comparisons. Between the six models, the minimum was 3611.05 (i.e., the Hurdle model), and the maximum was 4005.86 (i.e., the negative binomial Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle, ZIP, negative binomial Hurdle, negative binomial ZIP, Poisson, negative binomial Poisson. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-17 displays the descriptive statistics for these two models. Figure 4-19 is a boxplot illustrating these results.

The following set of results pertains to the data in which 250 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 750 values were sampled from a normal pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .25 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -69.34 to -5.87 with a mean of -35.75 and a standard deviation of 9.23. Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000 valid model comparisons. The average AIC for the Poisson model was 3832.24 while the average AIC for the negative binomial Poisson model was 3869.99, which descriptively supports the inferential findings of better fit for the Poisson model.

For the .25 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -19.83 to -3.93 with a mean of -5.2 and a standard deviation of .82. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,634 valid model comparisons. The average AIC for the Hurdle model was 3586.47 while the average AIC for the negative binomial Hurdle model was 3593.63 , which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .25 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 202.52 to 312.73 with a mean of 251.77 and a standard deviation of 14.35. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 3832.24 while the average AIC for the Hurdle model was 3586.47, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .25 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 228.78 to 330.2 with a mean of 282.08 and a standard deviation of 14.95. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,634 valid model comparisons. The average AIC for the negative binomial Poisson model was 3869.99 while the average AIC for the negative binomial Hurdle model was 3593.63, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .25 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -3.09 to -1.94 with a mean of -2.01 and a standard deviation of .05. Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000 valid model comparisons. The average AIC for the ZIP model was 3592.44 while the average AIC for the negative binomial ZIP model was 3596.45, which descriptively suggests better fit for the ZIP model.

Table 4-18 summarizes the log-likelihood statistics for the normal distribution model comparisons with a .25 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 3529.93 and a maximum of 2966.65 across all 11,634 valid model comparisons. Between the six models, the minimum was 3586.47 (i.e., the Hurdle model), and the maximum was 3869.99 (i.e., the negative binomial Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle, ZIP, negative binomial Hurdle, negative binomial ZIP, Poisson, negative binomial Poisson. These

rankings are not in line with the rankings for the pseudo-population AIC's; in the pseudo-population, the negative binomial Hurdle model has a lower AIC (i.e., better fit) than the ZIP model Table 4-19 displays the descriptive statistics for these two models. Figure 4-20 is a boxplot illustrating these results.

The following set of results pertains to the data in which 500 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 500 values were sampled from a normal pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .50 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 286.84 to 381.11 with a mean of 331.24 and a standard deviation of 13.74. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 3642.82 while the average AIC for the negative binomial Poisson model was 3356.12, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .50 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -41.05 to 3.15 with a mean of -3.86 and a standard deviation of 2.19 . Given a Type I error rate of .05, the deviance statistic was significant for none of the 1,602 valid model comparisons. The average AIC for the Hurdle model was 2978.57 while the average AIC for the negative binomial Hurdle model was 2994.57, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .50 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 655.17 to 857.57 with a mean of 758.23 and a standard deviation of 31.92. Given a Type I error rate of .05, the deviance statistic was significant all of the 2,000 valid model

comparisons. The average AIC for the Poisson model was 3642.82 while the average AIC for the Hurdle model was 3023.52, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .50 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 356.28 to 489.60 with a mean of 423.12 and a standard deviation of 21.81. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,601 valid model comparisons. The average AIC for the model was 3356.12 while the average AIC for the negative binomial Poisson model was 2994.57, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .50 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -2.57 to -1.93 with a mean of -2.01 and a standard deviation of .029 . Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000 valid model comparisons. The average AIC for the ZIP model was 2992.55 while the average AIC for the negative binomial ZIP model was 2996.55, which descriptively suggests better fit for the ZIP model.

Table 4-20 summarizes the log-likelihood statistics for the normal distribution model comparisons with a .50 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 2992.55 and a maximum of 3926.85 across all 12,000 valid model comparisons. Between the six models, the minimum

was 3023.52 (i.e., the Hurdle model), and the maximum was 3783.75 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle, ZIP, negative binomial Hurdle, negative binomial ZIP, negative binomial Poisson, Poisson. These rankings are not in line with the rankings for the pseudo-population AIC's; in the population, the negative binomial Hurdle model has a lower AIC (i.e., better fit) than the ZIP model. Table 4-21 displays the descriptive statistics for these two models. Figure 4-21 is a boxplot illustrating these results.

The following set of results pertains to the data in which 750 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 250 values were sampled from a normal pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .75 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 313.34 to 435.26 with a mean of 372.68 and a standard deviation of 17.63. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 2930.85 while the average AIC for the negative binomial Poisson model was 2561.17, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .75 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -36.75 to .10 with a mean of -2.29 and a standard deviation of .24. Given a Type I error rate of .05, the deviance statistic was significant for none of the 1,491 valid model comparisons. The average AIC for the Hurdle model was 1952.86 while the average AIC for the negative binomial Hurdle model was 1957.07, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .75 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 827.03 to 1132.06 with a mean of 983.00 and a standard deviation of 47.54. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 2930.85 while the average AIC for the Hurdle model was 1952.86, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .75 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 499.96 to 701.73 with a mean of 608.46 and a standard deviation of 30.96. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,490 valid model comparisons. The average AIC for the negative binomial Poisson model was 2560.17 while the average AIC for the negative binomial Hurdle model was 1957.07, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .75 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -5.86 to -.64 with a mean of -2.01 and a standard deviation of .10. Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000 valid model comparisons. The average AIC for the ZIP model was 1958.85 while the average AIC for the negative binomial ZIP model was 1962.87, which suggests adequate fit for the ZIP model.

Table 4-22 summarizes the log-likelihood statistics for the normal distribution model comparisons with a .75 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 1918.91 and a maximum of 3103.14 across all valid model comparisons. Between the six models, the minimum was 1952.86 (i.e., the Hurdle model), and the maximum was 2930.85 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle, negative binomial Hurdle, ZIP, negative binomial ZIP, negative binomial Poisson, Poisson. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-23 displays the descriptive statistics for these two models. Figure 4-22 is a boxplot illustrating these results.

The following set of results pertains to the data in which 900 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 100 values were sampled from a normal pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .90 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 157.17 to 249.62 with a mean of 204.06 and a standard deviation of 14.38. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 1716.21 while the average AIC for the negative binomial Poisson model was 1514.15, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .90 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -20.63 to 13.74 with a mean of -1.11 and a standard deviation of 1.62 . Given a Type I error rate of .05, the deviance statistic was significant for none of the 1,537 valid model comparisons. The average AIC for the Hurdle model was 984.24 while the average AIC for the negative binomial Hurdle model was 987.32, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .90 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 583.77 to 893.39 with a mean of 737.97 and a standard deviation of 49.17. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 1716.2 while the average AIC for the Hurdle model was 1514.15, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .90 proportion of zeros condition with a normal distribution, the deviance statistic ranged from 425.97 to 642.72 with a mean of 532.39 and a standard deviation of 34.87. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,537 valid model comparisons. The average AIC for the model negative binomial Poisson was 1514.15 while the average AIC for the negative binomial Hurdle model was 984.24, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .90 proportion of zeros condition with a normal distribution, the deviance statistic ranged from -2.53 to -1.97 with a mean of -2.01 and a standard deviation of .02. Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000 valid model comparisons. The average AIC for the ZIP model was 990.22 while the average AIC for the negative binomial ZIP model was 994.23 , which descriptively suggests better fit for the ZIP model.

Table 4-24 summarizes the log-likelihood statistics for the normal distribution model comparisons with a .90 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated

dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 962.46 and a maximum of 1881.02 across all 11,536 valid model comparisons. Between the six models, the minimum was 984.24 (i.e., the Hurdle model), and the maximum was 1716.20 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle, negative binomial Hurdle, ZIP, negative binomial ZIP, negative binomial Poisson, Poisson. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-25 displays the descriptive statistics for these two models. Figure 4-23 is a boxplot illustrating these results.

Negatively Skewed Distribution

The following set of results pertains to the data in which 100 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 900 values were sampled from a negatively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .10 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -572.46 to -470.93 with a mean of -517.56 and a standard deviation of 13.66. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 model comparisons. The average AIC for the Poisson model was 4044.50 while the average AIC for the negative binomial Poisson model was 4563.58, which descriptively supports the inferential findings of better fit for the Poisson model.

For the .10 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 13.4 to 61.02 with a mean of 15.75 and a standard deviation of .28. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,626

valid model comparisons. The average AIC for the Hurdle model was 3868.06 while the average AIC for the negative binomial Hurdle model was 3886.31, which descriptively support the inferential findings of better fit for the Hurdle model.

For the .10 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 162.77 to 202.59 with a mean of 181.96 and a standard deviation of 5.72. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 model comparisons. The average AIC for the Poisson model was 4044.02 while the average AIC for the Hurdle model was 3868.06, which descriptively support the inferential findings of better fit for the Hurdle model.

For the .10 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 609.07 to 754.25 with a mean of 682.91 and a standard deviation of 17.72. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,626 valid model comparisons. The average AIC for the negative binomial Poisson model was 4563.58 while the average AIC for the negative binomial Hurdle model was 3886.31, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .10 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -2.49 to -1.92 with a mean of -2.01 and a standard deviation of 0.03. Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000 valid model comparisons. The average AIC for the ZIP model was 3874.07 while the average AIC for the negative binomial Hurdle model was 3878.08, which descriptively better fit for the ZIP model.

Table 4-26 summarizes the log-likelihood statistics for the negatively skewed distribution model comparisons with a .10 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 3819.85 and a maximum of 4598.21 across all 9,626 valid model comparisons. Between the six models, the minimum was 3868.06 (i.e., the Hurdle model), and the maximum was 4563.58 (i.e., the negative binomial Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle model, ZIP model, negative binomial ZIP model, negative binomial Hurdle model, Poisson model, and negative binomial Poisson model. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-27 displays the descriptive statistics for these two models. Figure 4-24 is a boxplot illustrating these results.

The following set of results pertains to the data in which 250 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 750 values were sampled from a negatively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .25 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 30.72 to 86.61 with a mean of 57.02 and a standard deviation of .99. Given a Type I error rate of .05, the deviance statistic was significant for of the all 2,000 model comparisons. The average AIC for the Poisson model was 4534.49 while the average AIC

for the negative binomial Poisson model was 4479.47, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .25 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -48.44 to -10.95 with a mean of -13.01 and a standard deviation of 1.47. Given a Type I error rate of .05, the deviance statistic was significant for of the all 1,517 valid model comparisons. The average AIC for the Hurdle model was 3807.17 while the average AIC for the negative binomial Hurdle model was 3822.26 which descriptively the inferential findings of better fit for the Hurdle model.

For the .25 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 674.99 to 784.98 with a mean of 733.31 and a standard deviation of 15.50. Given a Type I error rate of .05, the deviance statistic was significant for of the all 2,000 valid model comparisons. The average AIC for the Poisson model was 4534.49, while the average AIC for the Hurdle model was 3807.17 which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .25 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 602.99 to 723.43 with a mean of 663.28 and a standard deviation of 19.81. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,517 valid model comparisons. The average AIC for the negative binomial Poisson model was 4479.47 while the average AIC for the negative binomial Hurdle model was 3822.26, which descriptively the inferential findings of better fit for the negative binomial Hurdle model.

For the .25 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -4.51 to -2.01 with a mean of -1.95 and a standard deviation of .06. Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000

valid model comparisons. The average AIC for the ZIP model was 3813.16 while the average AIC for the negative binomial ZIP model was 3817.17, which descriptively supports better fit for the ZIP model.

Table 4-28 summarizes the log-likelihood statistics for the negatively skewed distribution model comparisons with a .25 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 3765.15 and a maximum of 4576.05 across all 11,517 valid model comparisons. Between the six models, the minimum was 3,807.17 (i.e., the Hurdle model), and the maximum was 4534.49 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle model, ZIP model, negative binomial ZIP model, negative binomial Hurdle model, negative binomial Poisson model, and Poisson model . These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-29 displays the descriptive statistics for these two models. Figure 4-25 is a boxplot illustrating these results.

The following set of results pertains to the data in which 500 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 500 values were sampled from a negatively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .50 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 618.37 to 694.3 with a mean of 658.45 and a standard deviation of

9.97. Given a Type I error rate of .05, the deviance statistic was significant for all 2,000 valid model comparisons. The average AIC for the model Poisson model was 4747.63 while the average AIC for the negative binomial Poisson model was 4091.18, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .50 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -4.36 to 57.05 with a mean of 8.65 and a standard deviation of 2.17. Given a Type I error rate of .05, the deviance statistic was significant for 1,665 of the 1,672 valid model comparisons. The average AIC for the Hurdle model was 3179.05 while the average AIC for the negative binomial Hurdle model was 3189.83 which descriptively the 99.58% inferential findings of better fit for the Hurdle model.

For the .50 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 1472.82 to 1677.02 with a mean of 1574.58 and a standard deviation of 31.48. Given a Type I error rate of .05, the deviance statistic was significant for of all 2,000 valid model comparisons. The average AIC for the Poisson model was 4747.63 while the average AIC for the Hurdle model was 3179.05 which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .50 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 823.6 to 984.79 with a mean of 907.16 and a standard deviation of 24.35. Given a Type I error rate of .05, the deviance statistic was significant for all 1,672 valid model comparisons. The average AIC for the negative binomial Poisson model was 4091.18 while the average AIC for the negative binomial Hurdle model was 3189.83, which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .50 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -2.36 to -1.87 with a mean of -2.00 and a standard deviation of .02. Given a Type I error rate of .05, the deviance statistic was significant for none of the 1,998 valid model comparisons. The average AIC for the ZIP model was 3184.98 while the average AIC for the negative binomial ZIP model was 3180.98, which descriptively supports better fit for the negative binomial ZIP model.

Table 4-30 summarizes the log-likelihood statistics for the negatively skewed distribution model comparisons with a .50 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 3145.34 and a maximum of 4850.32 across all valid model comparisons. Between the six models, the minimum was 3179.05 (i.e., the Hurdle model), and the maximum was 4747.63 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle model, negative binomial ZIP model, ZIP model, negative binomial Hurdle model, negative binomial Poisson model, and Poisson model. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-31 displays the descriptive statistics for these two models. Figure 4-26 is a boxplot illustrating these results.

The following set of results pertains to the data in which 750 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 250 values were sampled

from a negatively skewed pseudo-population with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .75 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 647.7 to 760.47 with a mean of 704.52 and a standard deviation of 16.27. Given a Type I error rate of .05, the deviance statistic was significant for all 2,000 valid model comparisons. The average AIC for the Poisson model was 3845.4 while the average AIC for the negative binomial Poisson model was 3142.87, which descriptively supports the inferential findings of better fit for the negative binomial Poisson model.

For the .75 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -171.64 to 9.79 with a mean of -4.89 and a standard deviation of 5.61. Given a Type I error rate of .05, the deviance statistic was significant for 1418 of the 1563 valid model comparisons. The average AIC for the Hurdle model was 2027.47 while the average AIC for the negative binomial Hurdle model was 2034.39, which descriptively supports the 90.72% inferential findings of better fit for the Hurdle model.

For the .75 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 1472.82 to 1677.82 with a mean of 1574.58 and a standard deviation of 31.48. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 3845.4 while the average AIC for the Hurdle model was 2027.47 which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .75 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 976.32 to 1218.4 with a mean of 1114.55 and a standard deviation of 31.27. Given a Type I error rate of .05, the deviance statistic was significant for all of the

1,562 valid model comparisons. The average AIC for the negative binomial Poisson model was 3142.87 while the average AIC for the negative binomial Hurdle model was 2034.39 which descriptively supports the findings of better fit for the negative binomial Hurdle model.

For the .75 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -8.54 to -1.59 with a mean of -2.01 and a standard deviation of .15. Given a Type I error rate of .05, the deviance statistic was significant for 1of the 1,999 valid model comparisons. The average AIC for the ZIP model was 2033.45 while the average AIC for the negative binomial ZIP model was 2037.46, which descriptively supports the one inferential finding of better fit for the ZIP model.

Table 4-32 summarizes the log-likelihood statistics for the negatively skewed distribution model comparisons with a .75 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 2000.2 and a maximum of 3994.94 across all 11,561 valid model comparisons. Between the six models, the minimum was 2027.47 (i.e., the Hurdle model), and the maximum was 3845.4 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle model, ZIP model, negative binomial Hurdle model, negative binomial ZIP model, negative binomial Poisson model, and Poisson model. These rankings are, for the most part, in line with the rankings for the pseudo-population AIC's. However, in the pseudo-population, the AIC's for the Hurdle model and the

negative binomial Hurdle model were tied at 2020.3. Table 4-33 displays the descriptive statistics for these two models. Figure 4-27 is a boxplot illustrating these results.

The following set of results pertains to the data in which 900 of the 1,000 observations in each of the 2,000 simulated datasets were fixed at zero. The remaining 100 values were sampled from a negatively skewed pseudo-population distribution with the count proportions and frequencies displayed in Table 3-1 and Table 3-2.

For the .90 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 334.64 to 438.18 with a mean of 389.67 and a standard deviation of 14.71 . Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 2311.08, while the average AIC for the negative binomial Poisson model was 1923.4, which descriptively the inferential findings of better fit for the negative binomial Poisson model.

For the .90 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -40.33 to -1.06 with a mean of -2.03 and a standard deviation of 2.11 . Given a Type I error rate of .05, the deviance statistic was significant for 57 of the 1,478 valid model comparisons. The average AIC for the Hurdle model was 1017.47, while the average AIC for the negative binomial Hurdle model was 1021.39 which descriptively supports findings of better fit for the Hurdle model; inferentially, better fit was found for the negative binomial Hurdle model for 3.9% of the simulations.

For the .90 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 1119.3 to 1443.11 with a mean of 1299.61 and a standard deviation of 48.19. Given a Type I error rate of .05, the deviance statistic was significant for all of the 2,000 valid model comparisons. The average AIC for the Poisson model was 2311.08

while the average AIC for the Hurdle model was 1017.47, which descriptively supports the inferential findings of better fit for the Hurdle model.

For the .90 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from 783.55 to 997.99 with a mean of 907.75 and a standard deviation of 33.26. Given a Type I error rate of .05, the deviance statistic was significant for all of the 1,478 valid model comparisons. The average AIC for the negative binomial Poisson model was 1923.40 while the average AIC for the negative binomial Hurdle model was 1021.39 which descriptively supports the inferential findings of better fit for the negative binomial Hurdle model.

For the .90 proportion of zeros condition with a negatively skewed distribution, the deviance statistic ranged from -3.33 to -1.97 with a mean of -2.01 and a standard deviation of .04. Given a Type I error rate of .05, the deviance statistic was significant for none of the 2,000 valid model comparisons. The average AIC for the ZIP model was 1023.46 while the average AIC for the negative binomial ZIP model was 1027.47, which descriptively supports better fit for the ZIP model.

Table 4-34 summarizes the log-likelihood statistics for the negatively skewed distribution model comparisons with a .90 proportion of zeros. In this table, μ_{LL} represents the mean log-likelihood for the pseudo-population, $\overline{LL}$ represents the mean log-likelihood for the simulated dataset samples, $LL_{\min}$ and $LL_{\max}$ represent the log-likelihood minimum and maximum for the simulated dataset sample, s_{LL} and $s_{\overline{LL}}$ represent the standard deviation and standard error for the simulated dataset samples, and *n* represents the sample size for the simulated dataset samples.

A comparison of AIC's over all models revealed a minimum of 997.99 and a maximum of 2454.31 across all 11,478 valid model comparisons. Between the six models, the minimum was

1017.47 (i.e., the Hurdle model), and the maximum was 2311.08 (i.e., the Poisson model). The rank order from lowest AIC to highest AIC was as follows: Hurdle model, negative binomial Hurdle model, negative binomial Poisson, ZIP model, negative binomial ZIP model, and Poisson model. These rankings are in line with the rankings for the pseudo-population AIC's. Table 4-35 displays the descriptive statistics for these two models. Figure 4-28 is a boxplot illustrating these results.

Review of Positively Skewed Distribution Findings

For the .50, .75, and .90 proportions of zeros conditions, the negative binomial Poisson model displayed significantly better fit than the Poisson model; however, for the .10 and .25 proportions of zeros conditions, the Poisson model fit significantly better than the negative binomial Poisson model. For all conditions, the Hurdle model displayed significantly better fit than the Poisson model except at the .25 proportion of zeros condition where the Hurdle model displayed better fit for 25.2% of the simulations. The negative binomial Hurdle model displayed significantly better fit than the negative binomial Poisson model for all proportions zeros conditions and all simulations except for the .50 proportion of zeros condition in which the negative binomial Hurdle displayed significantly better fit for .133 of the simulations.

There was considerable variability in results comparing the Hurdle model and negative binomial Hurdle model as well as results comparing the ZIP model and negative binomial ZIP model. The negative binomial Hurdle displayed significantly better fit for all simulations when the proportion of zeros was .10 and .50. The proportion of simulations favoring the negative binomial Hurdle model dropped to .831 when the proportion of zeros of .75. This proportion significant dropped further to .307 for the .90 proportion of zeros conditions and dropped further yet to .252 for the .25 proportion of zeros conditions.

There were no conditions in which the negative binomial ZIP model displayed significantly better fit than the ZIP model. The proportion of such cases was .947 at the .50 proportion of zeros condition. The proportion dropped to .652 at the .75 proportion of zeros. There were only .133 and .186 simulations favoring the negative binomial ZIP model at .25 and .90 zeros, respectively. Finally, there were no simulations favoring the negative binomial ZIP model when the proportion of zeros was .10. However, as displayed in Table 4-42, there were convergence problems when the proportion of zeros was .10 and .25. Table 4-36 displays the percentage of simulations favoring the more complex model.

In terms of AIC, the results were similarly inconsistent. The negative binomial Hurdle model typically displayed the best fit; however, the negative binomial Poisson model, which displayed the worst fit for the .10 and .25 proportion of zeros conditions then displayed best fit for the .50 proportion of zeros condition. The Hurdle model displayed the second best fit when the proportion of zeros was .10. However, its rank decreased over subsequent simulations from third to fourth best-fitting model. The Poisson model also displayed inconsistent findings. For the .10 proportion of zeros condition, this model was the third best-fitting model. It then improved in fit at the .25 proportion of zeros condition but then became the worst fitting model for the remaining proportion of zeros conditions.

The ZIP model displayed fourth best fit when the proportion of zeros was .10 and .25; this rank was superseded by the negative binomial ZIP model. However, the results reversed for the .50 and .75 proportion of zeros conditions and reversed again at the .90 proportion of zeros conditions. It is important to note that these results are based on AIC's; inspection of previous AIC figures suggests that the rank order is purely for comparison since the actual AIC values may be nearly equal between models and/or proportion of zeros conditions. For example, the

superiority of the Hurdle model over the ZIP model .25 zeros is based on an AIC difference of 3.22 while the superiority of the Hurdle model over the ZIP model at .10 zeros is based on an AID difference of 129.04. Table 4-37 presents the AIC's for the positively skewed distribution over all models and proportions of zeros. Figure 4-29 illustrates the rank-ordering of these AIC's.

Review of Normal Distribution Findings

The results for the normally distributed event stage were consistent for all proportion of zeros. For all proportions, the Hurdle model displayed the best fit. There were no significant findings of support for the negative binomial Hurdle model. The two models did become more similar in terms of AIC's at the .75 and .90 proportions of zeros. This same pattern was found for the ZIP models. However, it is important to note that at the .25 and higher proportion of zeros conditions, Hurdle model comparisons converged for approximately 75% of the simulations while the ZIP model comparisons converged for 100% of the model comparisons. The convergence frequencies are displayed in Table 4-43. Both inferentially and descriptively, the Poisson model was superior to the negative binomial Poisson model when the proportion of zeros was .10 or .25; the reverse was true when the proportion of zeros was .50, .75, or .90. Table 4-38 displays the percentage of simulations significantly favoring the more complex model.

It is important to note that these results are based on AIC's; inspection of previous AIC figures suggests that the rank order is purely for comparison since the actual AIC values may be nearly equal between models and/or proportion of zeros conditions. For example, for all proportions of zeros, the difference in AIC's between the negative binomial Hurdle, ZIP, and ZIP models are quite small. Table 4-39 presents the AIC's for the normal distribution over all models and proportions of zeros. Figure 4-30 illustrates the rank-ordering over conditions.

Review of Negatively Skewed Distribution Findings

For all conditions except for the .10 proportion of zeros, the negative binomial Poisson model displayed significantly better fit than the Poisson model. Between all zero proportion conditions, the Hurdle model consistently displayed significantly better fit when compared to the Poisson model while the negative binomial Hurdle model consistently displayed significantly better fit when compared to the negative binomial Poisson model.

In general, the Hurdle model displayed significantly better fit than the negative binomial Hurdle model; however, the results should be interpreted cautiously since for all conditions, the negative binomial Hurdle had convergence problems of roughly 25%. This is displayed in Table 4-44. The negative binomial ZIP model never displayed significantly better fit than the ZIP model. Table 4-40 displays the percentage of simulations significantly favoring the more complex model.

Descriptively, the Hurdle model displayed the best fit regardless of the proportion of zeros. The Poisson model displayed the worst fit except for the .10 proportion of zeros condition in which the negative binomial Poisson model displayed the worst fit. Consequently, the next worst fitting model was this negative binomial Poisson model except for the .10 proportion of zeros. In between the best fitting and two worst fitting models, for the .10 and .25 proportions of zeros conditions, the ZIP model fit better than the negative binomial ZIP model, which fit better than the negative binomial Hurdle model. For the .50 proportion of zeros condition, the rank order for the ZIP and negative binomial ZIP was reversed with the negative binomial ZIP model displaying better fit. However, for the .75 proportion of zeros condition the ZIP model returned to being the second best-fitting model. Finally, for the .90 proportion of zeros condition the negative binomial Hurdle model gained in rank again to become the second best-fitting model.

The negative binomial Poisson also rose in rank to be the third best-fitting model. This was followed by the ZIP and negative binomial ZIP models.

It is important to note that these results are based on AIC's; inspection of previous AIC figures suggests that the rank order is purely for comparison since the actual AIC values may be nearly equal between models and/or proportion of zeros conditions. For example, although the AIC is lower for the Hurdle models than for the negative binomial Hurdle models, the difference between the AIC's declines sharply as the proportion of zeros increases. Table 4-41 presents the AIC's for the negatively skewed distribution over all models and proportions of zeros. Figure 4-31 illustrates the rank-ordering over conditions.

Table 4-1. Deviance statistics comparing Poisson and negative binomial Poisson models

	Distribution		
Zeros	Positive	Normal	Negative
.10	-238.40	-384.40	-517.40
.25	-82.60	-35.20	-1534.80
.50	88.40	331.00	658.40
.75	125.00	372.40	704.60
.90	69.80	203.00	389.60

Table 4-2. Deviance statistics comparing Hurdle and negative binomial Hurdle models

	Distribution		
Zeros	Positive	Normal	Negative
.10	27.60	-6.20	-15.60
.25	20.40	-5.20	-13.00
.50	15.40	-3.60	-8.40
.75	8.20	-1.80	-4.40
.90	3.20	-1.80	-1.80

Table 4-3. Deviance statistics comparing Poisson and Hurdle models

	Distribution		
Zeros	Positive	Normal	Negative
.10	131.60	19.20	181.80
.25	5.20	252.60	859.60
.50	75.40	756.80	1573.20
.75	215.40	983.00	1823.20
.90	203.20	732.40	1298.20

Table 4-4. Deviance statistics comparing NB Poisson and NB Hurdle models

	Distribution		
Zeros	Positive	Normal	Negative
.10	397.60	397.40	683.60
.25	108.20	282.60	662.20
.50	2.40	422.20	906.40
.75	98.60	608.80	1114.20
.90	136.60	527.60	906.80

Table 4-5. Deviance statistics comparing ZIP and negative binomial ZIP models

	Distribution		
Zeros	Positive	Normal	Negative
.10	-2.00	-2.00	-2.00
.25	-0.60	-2.00	-2.00
.50	14.20	-2.00	-2.00
.75	6.20	-2.00	-2.00
.90	1.20	-2.00	-2.00

Table 4-6. Log-likelihood comparisons for positively skewed distribution with .10 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1475.3	-1475.53	-1539.15	-1409.23	19.27	0.430995	2000
	NB Poisson	-1594.5	-1594.65	-1650.38	-1539.67	15.79	0.353131	2000
2	Hurdle	-1409.5	-1409.40	-1495.97	-1315.41	26.27	0.587599	2000
	NB Hurdle	-1395.7	-1395.44	-1483.98	-1304.26	25.68	0.574254	2000
3	Poisson	-1475.3	-1475.53	-1539.15	-1409.23	19.27	0.430995	2000
	Hurdle	-1409.5	-1409.40	-1495.97	-1315.41	26.27	0.587599	2000
4	NB Poisson	-1594.5	-1594.65	-1650.38	-1539.67	15.79	0.353131	2000
	NB Hurdle	-1395.7	-1395.44	-1483.98	-1304.26	25.68	0.574254	2000
5	ZIP	-1478.3	-1477.92	-1542.15	-1412.24	19.8380	0.505194	1542
	NB ZIP	-1479.3	-1485.34	-1531.90	-1439.61	18.73000	1.570000	136

Table 4-7. AIC's for positively skewed distribution models with a .10 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	2956.6	2957.06	2824.47	3084.30	38.55	0.86	2000
NB Poisson	3194.9	3193.30	3083.34	3304.77	31.58	0.71	2000
Hurdle	2822.9	2838.81	2650.83	3011.94	52.56	1.18	2000
NB Hurdle	2975.4	2804.88	2622.53	2981.96	51.36	1.15	2000
ZIP	2962.6	2967.85	2836.47	3096.30	39.68	1.01	1542
NB ZIP	2964.6	2984.79	3077.79	2984.69	36.55	3.13	136

Table 4-8. Log-likelihood comparisons for positively skewed distribution with .25 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1468.1	-1466.83	-1542.31	-1388.77	22.76	0.509050	2000
	NB Poisson	-1509.4	-1508.43	-1561.66	-1448.08	17.54	0.392323	2000
2	Hurdle	-1465.5	-1463.95	-1542.28	-1377.41	24.29	0.543358	2000
	NB Hurdle	-1455.3	-1453.50	-1523.70	-1370.03	23.67	0.529329	2000
3	Poisson	-1468.1	-1466.83	-1542.31	-1388.77	22.76	0.509050	2000
	Hurdle	-1465.5	-1463.95	-1542.28	-1377.41	24.29	0.543358	2000
4	NB Poisson	-1509.4	-1508.43	-1561.66	-1448.08	17.54	0.392323	2000
	NB Hurdle	-1455.3	-1453.50	-1523.70	-1370.03	23.67	0.529329	2000
5	ZIP	-1471.1	-1469.56	-1544.96	-1391.77	22.71	0.511613	1972
	NB ZIP	-1471.4	-1471.07	-1537.28	-1407.36	20.36	0.470331	1874

Table 4-9. AIC's for positively skewed distribution models with a .25 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	2942.3	2939.654	2783.538	3090.621	45.53082	1.018100	2000
NB Poisson	3026.8	3020.851	2900.160	3127.329	35.09046	0.784647	2000
Hurdle	2943.0	2947.899	2774.825	3104.551	48.59943	1.086716	2000
NB Hurdle	2922.5	2921.001	2754.053	3061.392	47.34459	1.058657	2000
ZIP	2954.3	2951.124	2795.538	3101.917	45.43863	1.023227	1972
NB ZIP	2956.7	2956.147	2828.716	3088.553	40.72097	0.940661	1874

Table 4-10. Log-likelihood comparisons for positively skewed distribution with .50 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1334.3	-1333.44	-1428.47	-1239.52	29.11	0.650	2000
	NB Poisson	-1290.1	-1289.49	-1361.70	-1223.90	21.34	0.477	2000
2	Hurdle	-1296.6	-1294.96	-1357.10	-1221.90	20.80	0.465	2000
	NB Hurdle	-1288.9	-1287.12	-1352.33	-1220.42	20.26	0.906	2000
3	Poisson	-1334.3	-1333.44	-1428.47	-1239.52	29.11	0.650	2000
	Hurdle	-1296.6	-1294.96	-1357.10	-1221.90	20.80	0.465	2000
4	NBPoisson	-1290.1	-1289.49	-1361.70	-1223.90	21.34	0.477	2000
	NB Hurdle	-1288.9	-1287.12	-1352.33	-1220.42	20.26	0.906	2000
5	ZIP	-1299.5	-1297.92	-1360.09	-1224.63	20.80	0.465	2000
	NB ZIP	-1292.4	-1291.12	-1356.33	-1224.00	20.17	0.451	2000

Table 4-11. AIC's for positively skewed distribution models with a .50 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	2674.5	2672.876	2485.036	2862.935	58.22	1.301853	2000
NB Poisson	2588.3	2582.975	2451.801	2727.394	42.68	0.954555	2000
Hurdle	2605.1	2601.929	2455.800	2726.197	41.60	0.930422	2000
NBHurdle	2589.8	2588.240	2454.830	2718.660	40.52	0.905000	2000
ZIP	2611.1	2607.840	2461.262	2732.189	41.61	0.930524	2000
NB ZIP	2598.8	2596.232	2462.004	2726.654	40.35	0.902392	2000

Table 4-12. Log-likelihood comparisons for positively skewed distribution with .75 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-972.9	-974.608	-1088.690	-853.761	31.925	0.713	2000
	NB Poisson	-910.4	-911.681	-999.070	-819.438	23.872	0.533	2000
2	Hurdle	-865.2	-864.852	-907.808	-806.146	14.387	0.321	2000
	NB Hurdle	-861.1	-860.701	-905.917	-801.635	14.148	0.316	2000
3	Poisson	-972.9	-974.608	-1088.690	-853.761	31.925	0.713	2000
	Hurdle	-865.2	-864.852	-907.808	-806.146	14.387	0.321	2000
4	NB Poisson	-910.4	-911.681	-999.070	-819.438	23.872	0.533	2000
	NB Hurdle	-861.1	-860.701	-905.917	-801.635	14.148	0.316	2000
5	ZIP	-868.2	-867.858	-910.807	-809.105	14.382	0.321	2000
	NB ZIP	-865.1	-864.721	-909.922	-805.618	14.135	0.316	2000

Table 4-13. AIC's for positively skewed distribution models with a .75 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	1951.7	1955.215	1713.523	2183.381	63.85	1.42776	2000
NB Poisson	1828.7	1827.363	1642.875	2002.140	47.74	1.06762	2000
Hurdle	1742.5	1749.705	1632.292	1835.616	28.77	0.64343	2000
NB Hurdle	1734.3	1735.402	1617.270	1825.834	28.29	0.63273	2000
ZIP	1750.5	1747.716	1630.2100	1833.615	28.76	0.64320	2000
NB ZIP	1744.3	1743.442	1625.237	1833.844	28.27	0.63215	2000

Table 4-14. Log-likelihood comparisons for positively skewed distribution with .90 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-546.5	-546.397	-653.705	-453.190	29.33	0.655	2000
	NB Poisson	-511.6	-511.489	-594.277	-437.848	23.20	0.518	2000
2	Hurdle	-444.9	-443.180	-469.812	-408.241	9.020	0.201	2000
	NB Hurdle	-443.3	-441.750	-468.410	-410.220	8.848	0.198	1992
3	Poisson	-546.5	-546.397	-653.705	-453.190	29.33	0.655	2000
	Hurdle	-444.9	-443.180	-469.812	-408.241	9.020	0.201	2000
4	NB Poisson	-511.6	-511.489	-594.277	-437.848	23.20	0.518	2000
	NB Hurdle	-443.3	-441.750	-468.410	-410.220	8.848	0.198	1992
5	ZIP	-447.9	-446.170	-472.716	-411.142	9.021	0.201	2000
	NB ZIP	-447.3	-445.691	-472.314	-412.142	8.876	0.198	2000

Table 4-15. AIC's for positively skewed distribution models with a .90 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	1099.0	1098.794	912.398	1313.41	58.66007	1.311679	2000
NB Poisson	1031.2	1026.978	879.696	1192.55	46.40218	1.037584	2000
Hurdle	901.8	906.361	836.481	959.623	18.04078	0.403404	2000
NB Hurdle	898.7	897.499	834.439	950.819	17.69794	0.396532	1992
ZIP	907.7	904.339	834.283	957.432	18.04285	0.403450	2000
NB ZIP	908.7	905.381	838.283	958.627	17.75302	0.396970	2000

Table 4-16. Log-likelihood comparisons for normal distribution with .10 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1806.7	-1806.38	-1834.64	-1776.08	10.080	0.225	2000
	NB Poisson	-1998.9	-1998.93	-2031.97	-1961.23	9.706	0.217	2000
2	Hurdle	-1797.1	-1822.85	-1768.77	-1822.85	9.086	0.203	2000
	NB Hurdle	-1800.2	-1825.6	-1772.63	-1825.60	9.330	0.240	1503
3	Poisson	-1806.7	-1806.38	-1834.64	-1776.08	10.080	0.225	2000
	Hurdle	-1797.1	-1822.85	-1768.77	-1822.85	9.086	0.203	2000
4	NB Poisson	-1998.9	-1998.93	-2031.97	-1961.23	9.706	0.217	2000
	NB Hurdle	-1800.2	-1799.77	-1825.60	-1772.63	9.330	0.240	1503
5	ZIP	-1800.1	-1799.53	-1825.87	-1771.78	9.085	0.203	2000
	NB ZIP	-1801.1	-1800.63	-1826.87	-1772.78	9.042	0.203	1982

Table 4-17. AIC's for normal distribution models with a .10 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	3619.3	3618.760	3558.160	3675.274	20.17	0.45	2000
NB Poisson	4005.9	4005.863	3930.452	4071.946	19.41	0.43	2000
Hurdle	3606.1	3605.028	3549.532	3657.695	18.17	0.40	2000
NB Hurdle	3612.3	3613.546	3559.261	3665.195	18.66	0.48	1503
ZIP	3612.1	3611.051	3555.562	3663.733	18.17	0.40	2000
NB ZIP	3616.1	3615.268	3559.562	3667.739	18.08	0.40	1982

Table 4-18. Log-likelihood comparisons for normal distribution with .25 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1913.9	-1913.12	-1968.17	-1869.47	13.99	0.312	2000
	NB Poisson	-1931.5	-1930.99	-1979.33	-1889.82	11.77	0.263	2000
2	Hurdle	-1787.6	-1787.24	-1813.49	-1758.97	8.104	0.181	2000
	NB Hurdle	-1790.2	-1789.81	-1816.07	-1764.27	8.152	0.201	1634
3	Poisson	-1913.9	-1913.12	-1968.17	-1869.47	13.999	0.312	2000
	Hurdle	-1787.6	-1787.24	-1813.49	-1758.97	8.104	0.181	2000
4	NB Poisson	-1931.5	-1930.99	-1979.33	-1889.82	11.775	0.263	2000
	NB Hurdle	-1790.2	-1789.81	-1816.07	-1764.27	8.152	0.201	1634
5	ZIP	-1790.6	-1790.22	-1816.47	-1761.96	8.105	0.181	2000
	NB ZIP	-1791.6	-1791.22	-1817.47	-1762.97	8.105	0.181	2000

Table 4-19. AIC's for normal distribution models with a .25 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	3833.8	3832.239	3744.945	3942.340	27.99438	0.625973	2000
NB Poisson	3871.0	3869.986	3787.645	3966.654	23.55270	0.526654	2000
Hurdle	3587.2	3586.473	3529.931	3638.976	16.20934	0.362452	2000
NB Hurdle	3592.3	3593.629	3542.533	3646.150	16.30441	0.403347	1634
ZIP	3593.1	3592.436	3535.917	3644.938	16.21145	0.362499	2000
NB ZIP	3597.1	3596.448	3539.944	3648.939	16.21105	0.362490	2000

Table 4-20. Log-likelihood comparisons for normal distribution with .50 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1888.5	-1888.88	-1960.42	-1818.41	20.88	0.466	2000
	NB Poisson	-1723.0	-1723.25	-1769.87	-1674.06	14.60	0.326	2000
2	Hurdle	-1510.1	-1509.76	-1532.97	-1487.29	6.62	0.148	2000
	NB Hurdle	-1511.9	-1509.76	-1532.97	-1487.29	6.62	0.148	1602
3	Poisson	-1888.5	-1888.88	-1960.42	-1818.41	20.88	0.466	2000
	Hurdle	-1510.1	-1509.76	-1532.97	-1487.29	6.627	0.148	2000
4	NB Poisson	-1723.0	-1723.25	-1769.87	-1674.06	14.60	0.326	2000
	NB Hurdle	-1511.9	-1509.76	-1532.97	-1487.29	6.62	0.148	1602
5	ZIP	-1513.1	-1512.75	-1535.96	-1490.28	6.61	0.147	2000
	NB ZIP	-1514.1	-1513.75	-1536.96	-1491.28	6.61	0.147	2000

Table 4-21. AIC's for normal distribution models with a .50 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	3783.0	3783.75	3642.81	3926.846	41.76	0.93387	2000
NB Poisson	3454.0	3454.50	3356.11	3547.741	29.21	0.653052	2000
Hurdle	3023.2	3023.52	2978.57	3069.939	13.25	0.296204	2000
NB Hurdle	3035.8	3039.52	2994.57	3085.939	13.25	0.296204	1602
ZIP	3038.3	3037.49	2992.55	3083.928	13.23	0.295936	2000
NB ZIP	3042.3	3041.50	2996.55	3087.928	13.23	0.295941	2000

Table 4-22. Log-likelihood comparisons for normal distribution with .75 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1461.8	-1462.42	-1548.57	-1371.53	27.106	0.606	2000
	NB Poisson	-1275.6	-1276.08	-1332.19	-1214.35	18.507	0.413	2000
2	Hurdle	-970.3	-970.42	-987.29	-953.45	4.779	0.106	2000
	NB Hurdle	-971.2	-971.53	-992.37	-954.37	4.993	0.129	1490
3	Poisson	-1461.8	-1462.42	-1548.57	-1371.53	27.106	0.606	2000
	Hurdle	-970.3	-970.42	-987.29	-953.45	4.779	0.106	2000
4	NB Poisson	-1275.6	-1276.08	-1332.19	-1214.35	18.507	0.413	2000
	NB Hurdle	-971.2	-971.53	-992.37	-954.37	4.993	0.129	1490
5	ZIP	-990.3	-973.30	-956.46	-990.29	4.770	0.106	2000
	NB ZIP	-991.3	-974.30	-957.45	-991.29	4.770	0.106	2000

Table 4-23. AIC's for normal distribution models with a .75 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	2929.6	2930.846	2749.052	3103.141	54.21258	1.21223	2000
NB Poisson	2559.3	2560.165	2436.693	2672.373	37.0152	0.82768	2000
Hurdle	1952.6	1952.855	1918.906	1986.58	9.559016	0.21374	2000
NB Hurdle	1954.4	1957.073	1922.741	1998.756	9.987633	0.25874	1490
ZIP	1958.5	1958.852	1924.922	1992.58	9.557559	0.21371	2000
NB ZIP	1962.6	1962.866	1928.918	1996.581	9.559452	0.21375	2000

Table 4-24. Log-likelihood comparisons for normal distribution with .90 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-852.6	-855.101	-937.508	-767.641	26.52	0.593	1999
	NB Poisson	-751.1	-753.075	-812.698	-689.054	19.45	0.434	2000
2	Hurdle	-486.4	-486.118	-494.801	-475.228	2.89	0.064	2000
	NB Hurdle	-487.3	-486.659	-498.835	-475.636	2.91	0.074	1537
3	Poisson	-852.6	-855.101	-937.508	-767.641	26.52	0.593	1999
	Hurdle	-486.4	-486.118	-494.801	-475.228	2.89	0.064	2000
4	NB Poisson	-751.1	-753.075	-812.698	-689.054	19.45	0.434	2000
	NB Hurdle	-487.3	-486.659	-498.835	-475.636	2.91	0.074	1537
5	ZIP	-489.4	-489.111	-497.799	-478.242	2.89	0.064	2000
	NB ZIP	-490.4	-490.114	-498.799	-479.243	2.89	0.064	2000

Table 4-25. AIC's for normal distribution models with a .90 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	1711.1	1716.202	1541.281	1881.015	53.0534	1.18660	1999
NB Poisson	1510.2	1514.149	1386.108	1633.397	38.9068	0.86998	2000
Hurdle	984.8	984.2365	962.4563	1001.602	5.79776	0.12964	2000
NB Hurdle	986.6	987.3187	965.2728	1011.67	5.82673	0.14862	1537
ZIP	990.8	990.2214	968.4848	1007.597	5.79280	0.12953	2000
NB ZIP	994.8	994.2271	972.4854	1011.597	5.79265	0.12958	2000

Table 4-26. Log-likelihood comparisons for negatively skewed distribution with .10 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-2019.20	-2019.01	-2035.41	-2001.08	5.26	.12	2000
	NB Poisson	-2277.90	-2277.79	-2296.11	-2258.54	5.97	.13	2000
2	Hurdle	-1928.30	-1928.03	-1948.88	-1907.92	5.82	.13	2000
	NB Hurdle	-1936.10	-1936.16	-1965.98	-1918.98	5.64	.14	1626
3	Poisson	-2019.20	-2019.01	-2035.41	-2001.08	5.26	.12	2000
	Hurdle	-1928.30	-1928.03	-1948.88	-1907.92	5.82	.13	2000
4	NB Poisson	-2277.90	-2277.79	-2296.11	-2258.54	5.97	.13	2000
	NB Hurdle	-1936.10	-1936.16	-1965.98	-1918.98	5.64	.14	1626
5	ZIP	-1931.30	-1931.03	1951.87	-1910.92	5.82	.13	2000
	NB ZIP	-1932.30	-1932.04	1952.87	-1911.92	5.82	.13	2000

Table 4-27. AIC's for negatively skewed models with a .10 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	4044.50	4044.02	4008.15	4076.81	10.53	.24	2000
NB Poisson	4563.90	4563.58	4525.08	4600.21	11.94	.27	2000
Hurdle	3868.70	3868.06	3827.85	3909.75	11.25	.28	2000
NB Hurdle	3884.10	3886.31	3851.96	3945.97	11.25	.28	1626
ZIP	3874.60	3874.07	3833.84	3915.74	11.65	.26	2000
NB ZIP	3878.70	3878.08	3837.85	3919.75	11.65	.26	2000

Table 4-28. Log-likelihood comparisons for negatively skewed distribution with .25 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-2264.3	-2264.24	-2285.03	-2239.14	7.13	0.16	2000
	NB Poisson	-2235.5	-2235.74	-2258.49	-2208.87	7.18	0.16	2000
2	Hurdle	-1897.9	-1897.59	-1917.18	-1880.57	5.24	0.12	2000
	NB Hurdle	-1904.4	-1904.13	-1922.89	-1887.67	5.12	0.13	1517
3	Poisson	-2264.3	-2264.24	-2285.03	-2239.14	7.13	0.16	2000
	Hurdle	-1897.9	-1897.59	-1917.18	-1880.57	5.24	0.12	2000
4	NB Poisson	-2235.5	-2235.74	-2258.49	-2208.87	7.18	0.16	2000
	NB Hurdle	-1904.4	-1904.13	-1922.89	-1887.67	5.12	0.13	1517
5	ZIP	-1900.9	-1900.58	-1920.18	-1883.57	5.24	0.12	2000
	NB ZIP	-1901.9	-1901.58	-1921.18	-1884.57	5.24	0.12	2000

Table 4-29. AIC's for negatively skewed models with a .25 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	4534.7	4534.49	4484.28	4576.05	14.25	0.32	2000
NB Poisson	4479.0	4479.47	4425.75	4524.99	14.37	0.32	2000
Hurdle	3807.9	3807.17	3773.15	3844.37	10.47	0.23	2000
NB Hurdle	3820.8	3822.26	3789.34	3859.78	10.25	0.26	1517
ZIP	3813.9	3813.16	3783.14	3852.36	10.48	0.23	2000
NB ZIP	3817.9	3817.17	3783.15	3856.36	10.48	0.23	2000

Table 4-30. Log-likelihood comparisons for negatively skewed distribution with .50 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-2370.5	-2370.81	-2422.16	-2317.09	14.38	0.32	2000
	NB Poisson	-2041.3	-2041.59	-2075.01	-2007.90	9.95	0.22	2000
2	Hurdle	-1583.9	-1583.52	-1597.10	-1570.67	4.44	0.10	2000
	NB Hurdle	-1588.1	-1587.91	-1615.10	-1576.02	4.42	0.11	1672
3	Poisson	-2370.5	-2370.81	-2422.16	-2317.09	14.38	0.32	2000
	Hurdle	-1583.9	-1583.52	-1597.10	-1570.67	4.44	0.10	2000
4	NB Poisson	-2041.3	-2041.59	-2075.01	-2007.90	9.95	0.22	2000
	NB Hurdle	-1588.1	-1587.91	-1615.10	-1576.02	4.42	0.11	1672
5	ZIP	-1586.9	-1586.49	-1600.10	-1573.67	4.43	0.10	2000
	NB ZIP	-1587.9	-1587.49	-1601.10	-1574.67	4.44	0.10	1998

Table 4-31. AIC's for negatively skewed models with a .50 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	4746.9	4747.63	4640.17	4850.32	28.76	0.64	2000
NB Poisson	4090.7	4091.18	4023.80	4158.02	19.89	0.44	2000
Hurdle	3179.7	3179.05	3153.34	3206.21	8.88	0.20	1672
NB Hurdle	3188.2	3189.83	3166.04	3244.19	8.83	0.22	2000
ZIP	3185.7	3184.98	3156.34	3212.20	8.87	0.20	2000
NB ZIP	3181.7	3180.98	3155.34	3208.20	8.87	0.20	1998

Table 4-32. Log-likelihood comparisons for negatively skewed distribution with .75 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1919.7	-1919.70	-1994.47	-1845.46	21.89	0.49	2000
	NB Poisson	-1567.4	-1567.44	-1614.24	-1519.31	13.88	0.31	2000
2	Hurdle	-1008.1	-1007.74	-1019.17	-998.10	3.14	0.07	2000
	NB Hurdle	-1008.2	-1010.19	-1085.59	-1000.67	4.05	0.10	1562
3	Poisson	-1919.7	-1919.70	-1994.47	-1845.46	21.89	0.49	2000
	Hurdle	-1008.1	-1007.74	-1019.17	-998.10	3.14	0.07	2000
4	NB Poisson	-1567.4	-1567.44	-1614.24	-1519.31	13.88	0.31	2000
	NB Hurdle	-1008.2	-1010.19	-1085.59	-1000.67	4.05	0.10	1562
5	ZIP	-1011.2	-1010.73	-1022.17	-1001.10	3.13	0.07	2000
	NB ZIP	-1012.2	-1011.73	-1023.17	-1002.10	3.13	0.07	1999

Table 4-33. AIC's for negatively skewed models with a .75 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	3845.3	3845.40	3696.92	3994.94	43.78	0.98	2000
NB Poisson	3142.8	3142.87	3046.63	3236.48	27.77	0.62	2000
Hurdle	2028.3	2027.47	2008.20	2050.34	6.27	0.14	2000
NB Hurdle	2028.3	2034.39	2015.34	2185.19	8.11	0.21	1562
ZIP	2034.3	2033.45	2014.20	2056.34	6.26	0.14	2000
NB ZIP	2038.3	2037.46	2018.20	2060.34	6.27	0.14	1999

Table 4-34. Log-likelihood comparisons for negatively skewed distribution with .90 zeros

	Model	μ_{LL}	$\overline{LL}$	$LL_{\min}$	$LL_{\max}$	s_{LL}	$s_{\overline{LL}}$	n
1	Poisson	-1152.1	-1152.54	-1224.16	-1064.51	23.17	0.52	2000
	NB Poisson	-957.3	-957.70	-1005.07	-897.19	15.88	0.36	2000
2	Hurdle	-503.0	-502.73	-510.67	-496.99	2.09	0.05	2000
	NB Hurdle	-503.9	-503.70	-524.86	-498.07	2.31	0.06	1478
3	Poisson	-1152.1	-1152.54	-1224.16	-1064.51	23.17	0.52	2000
	Hurdle	-503.0	-502.73	-510.67	-496.99	2.09	0.05	2000
4	NB Poisson	-957.3	-957.70	-1005.07	-897.19	15.88	0.36	2000
	NB Hurdle	-503.9	-503.70	-524.86	-498.07	2.31	0.06	1478
5	ZIP	-506.0	-505.73	-513.67	-499.99	2.09	0.05	2000
	NB ZIP	-507.0	-506.74	-514.67	-501.00	2.09	0.05	2000

Table 4-35. AIC's for negatively skewed models with a .90 proportion of zeros

Model	μ_{AIC}	$\overline{AIC}$	$AIC_{\min}$	$AIC_{\max}$	s_{AIC}	$s_{\overline{AIC}}$	n
Poisson	2310.2	2311.08	2135.01	2454.31	46.33	1.04	2000
NB Poisson	1922.5	1923.40	1802.37	2018.13	31.77	0.71	2000
Hurdle	1018.1	1017.47	1015.99	1033.35	4.18	0.09	2000
NB Hurdle	1019.8	1021.39	1010.15	1063.71	4.61	0.12	1478
ZIP	1024.1	1023.46	1011.99	1039.34	4.18	0.09	2000
NB ZIP	1028.1	1027.47	1015.99	1043.34	4.18	0.09	2000

Table 4-36. Positively skewed distribution: percentage of simulations favoring complex model.

Simple vs. Complex	.10	.25	.50	.75	.90
P vs. NBP	0.0	0.0	100.0	100.0	100.0
H vs. NBH	100.0	99.7	98.0	98.1	30.7
P vs. H	100.0	25.2	100.0	100.0	100.0
NBP vs. NBH	100.0	100.0	100.0	100.0	100.0
Z vs. NBZ	0.0	18.6	94.7	65.2	13.3

Table 4-37. AIC's: positively skewed distribution (all conditions)

	Proportion of Zeros				
Model	0.10	0.25	0.50	0.75	0.90
Poisson	2957.06	2939.65	2672.88	1955.22	1098.79
NB Poisson	3193.30	3020.85	2582.98	1827.36	1026.98
Hurdle	2838.81	2947.90	2601.93	1749.71	906.36
NB Hurdle	2804.88	2921.00	2588.24	1735.40	897.50
ZIP	2967.85	2951.12	2607.84	1747.72	904.34
NB ZIP	2984.79	2956.15	2596.23	1743.44	905.38

Table 4-38. Normal distribution: percentage of simulations favoring complex model.

Simple vs. Complex	.10	.25	.50	.75	.90
P vs. NB P	0.0	0.0	100.0	100.0	100.0
H vs. NBH	0.0	0.0	0.0	0.0	0.0
P vs. H	100.0	100.0	100.0	100.0	100.0
NBP vs. NBH	100.0	100.0	100.0	100.0	100.0
Z vs. NBZ	.4	0.0	0.0	0.0	0.0

Table 4-39. AIC's: normal distribution (all conditions)

	Proportion of Zeros				
Model	0.10	0.25	0.50	0.75	0.90
Poisson	3618.76	3832.24	3783.75	2930.85	1716.20
NB Poisson	4005.86	3869.99	3454.51	2560.17	1514.15
Hurdle	3605.03	3586.47	3023.52	1952.86	984.24
NB Hurdle	3613.55	3593.63	3039.52	1957.07	987.32
ZIP	3611.05	3592.44	3037.49	1958.85	990.22
NB ZIP	3615.27	3596.45	3041.50	1962.87	994.23

Table 4-40. Negatively skewed distribution: percentage of simulations favoring complex model.

Simple vs. Complex	.10	.25	.50	.75	.90
P vs. NB P	0.0	100.0	100.0	100.0	100.0
H vs. NBH	100.0	100.0	99.9	90.7	100.0
P vs. H	100.0	100.0	100.0	100.0	100.0
NBP vs. NBH	100.0	100.0	100.0	100.0	100.0
Z vs. NBZ	0.0	0.0	0.0	0.0	0.0

Table 4-41. AIC's: negatively skewed distribution (all conditions)

	Proportion of Zeros				
Model	0.10	0.25	0.50	0.75	0.90
Poisson	4044.02	4534.49	4747.63	3845.40	2311.08
NB Poisson	4563.58	4479.47	4091.18	3142.87	1923.40
Hurdle	3868.06	3807.17	3179.05	2027.47	1017.47
NB Hurdle	3886.31	3822.26	3189.83	2034.39	1021.39
ZIP	3874.07	3813.16	3184.98	2033.45	1023.46
NB ZIP	3878.08	3817.17	3180.98	2037.46	1027.47

Table 4-42. Convergence frequencies: positively skewed distribution

	Proportion of Zeros				
Model	0.10	0.25	0.50	0.75	0.90
Poisson	2000	2000	2000	2000	2000
NB Poisson	2000	2000	2000	2000	2000
Hurdle	2000	2000	2000	2000	2000
NB Hurdle	2000	2000	2000	2000	2000
ZIP	1542	1972	2000	2000	2000
NB ZIP	136	1874	2000	2000	1992

Table 4-43. Convergence frequencies: normal distribution

	Proportion of Zeros				
Model	0.10	0.25	0.50	0.75	0.90
Poisson	2000	2000	2000	2000	2000
NB Poisson	2000	2000	2000	2000	2000
Hurdle	1503	2000	2000	2000	2000
NB Hurdle	1982	1634	1602	1490	1537
ZIP	2000	2000	2000	2000	2000
NB ZIP	2000	2000	2000	2000	2000

Table 4-44. Convergence frequencies: negatively skewed distribution

	Proportion of Zeros				
Model	0.10	0.25	0.50	0.75	0.90
Poisson	2000	2000	2000	2000	2000
NB Poisson	2000	2000	2000	2000	2000
Hurdle	2000	2000	2000	2000	2000
NB Hurdle	1627	1517	1672	1562	1478
ZIP	2000	2000	2000	2000	2000
NB ZIP	2000	2000	2000	1999	2000

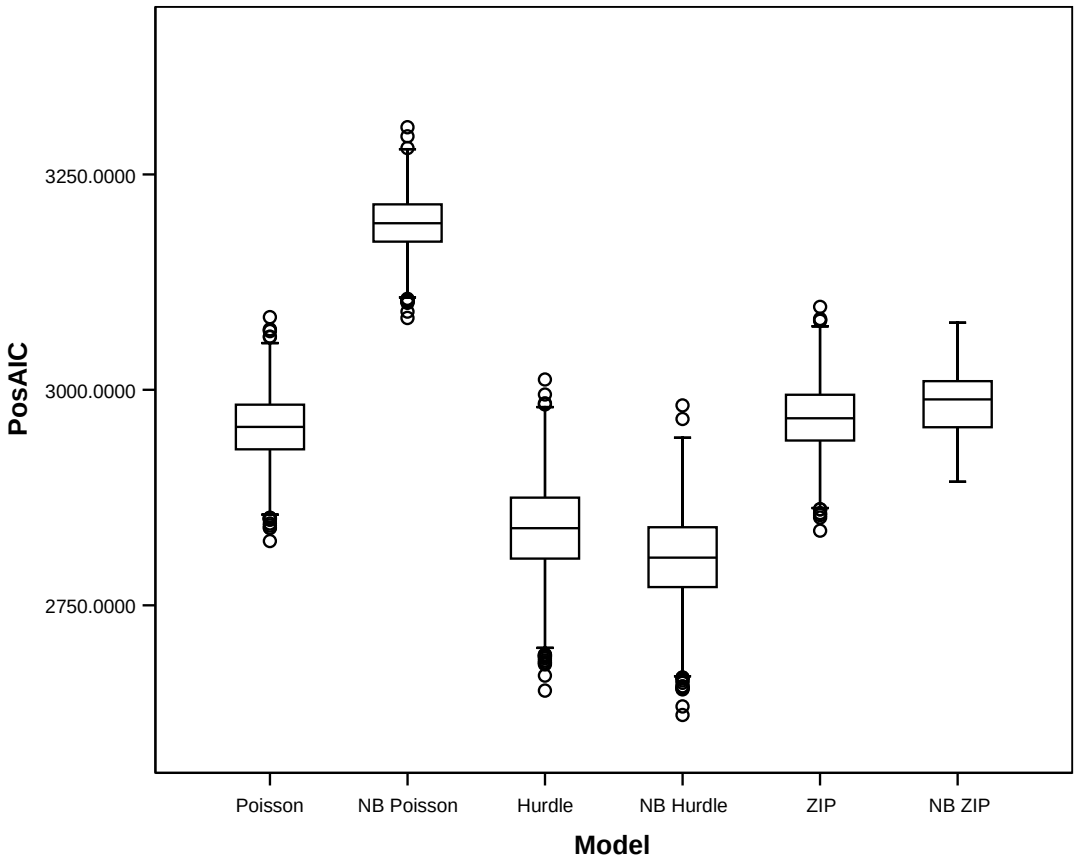

Figure 4-1. Boxplot of AIC's for all models for a .10 proportion of zeros

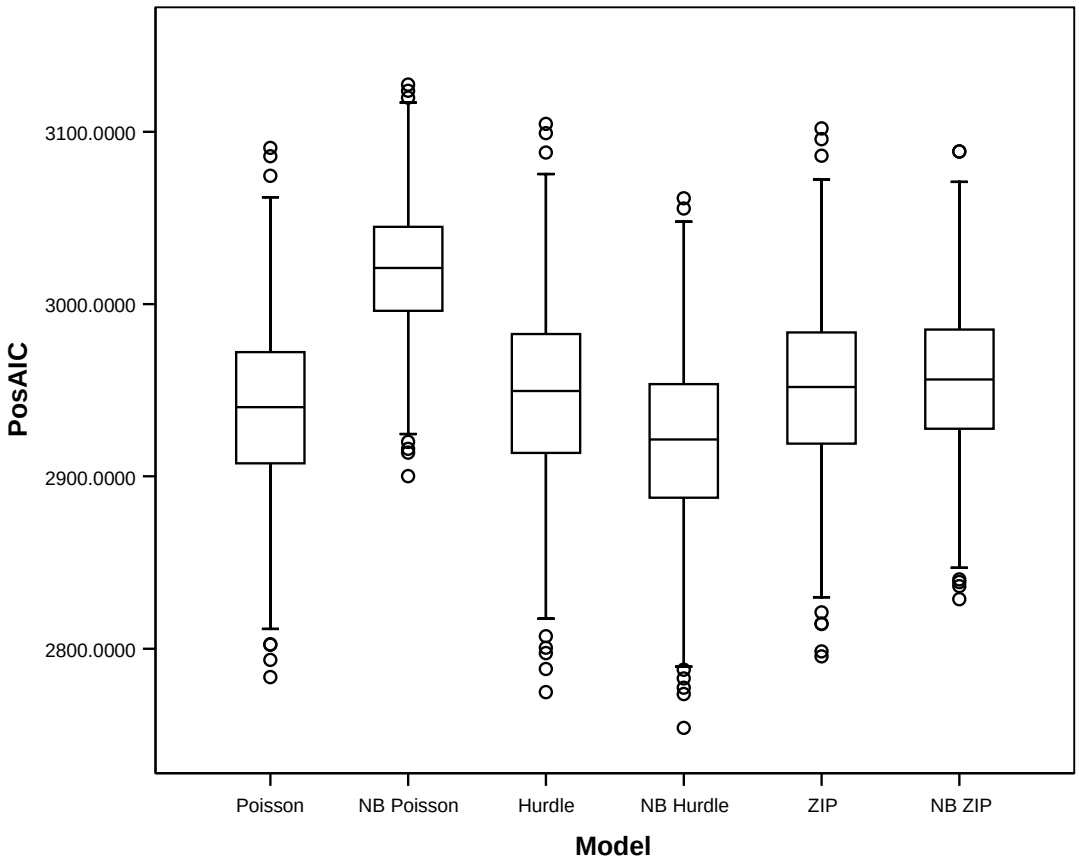

Figure 4-2. Boxplot of AIC's for all models for a .25 proportion of zeros

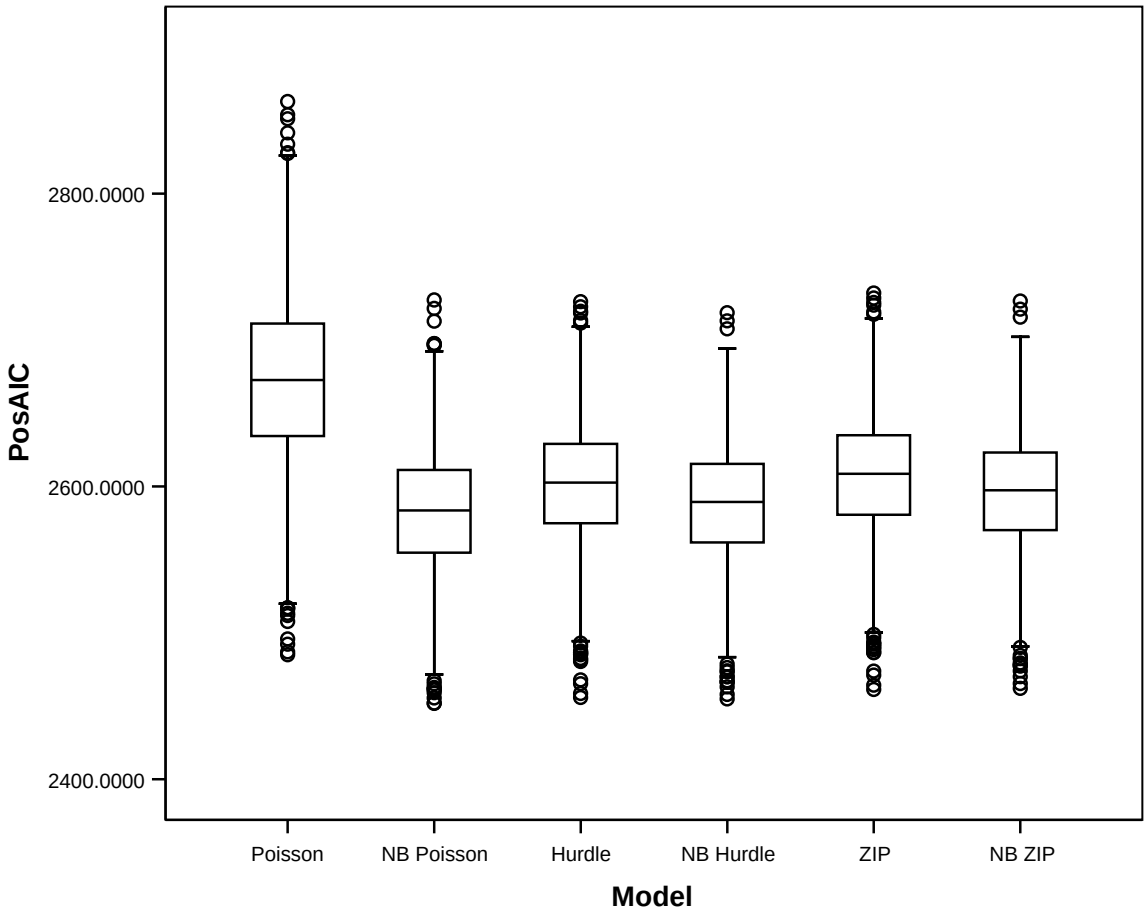

Figure 4-3. Boxplot of AIC's for all models for a .50 proportion of zeros

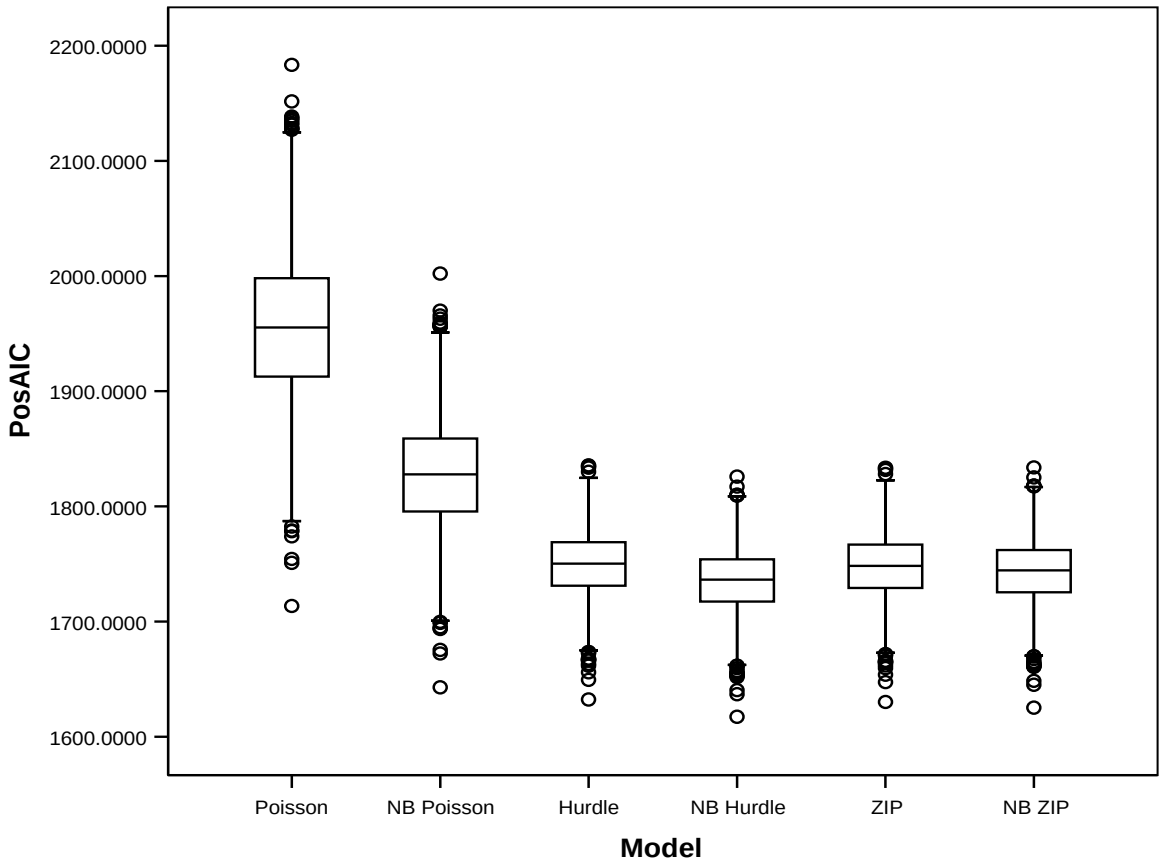

Figure 4-4. Boxplot of AIC's for all models for a .75 proportion of zeros

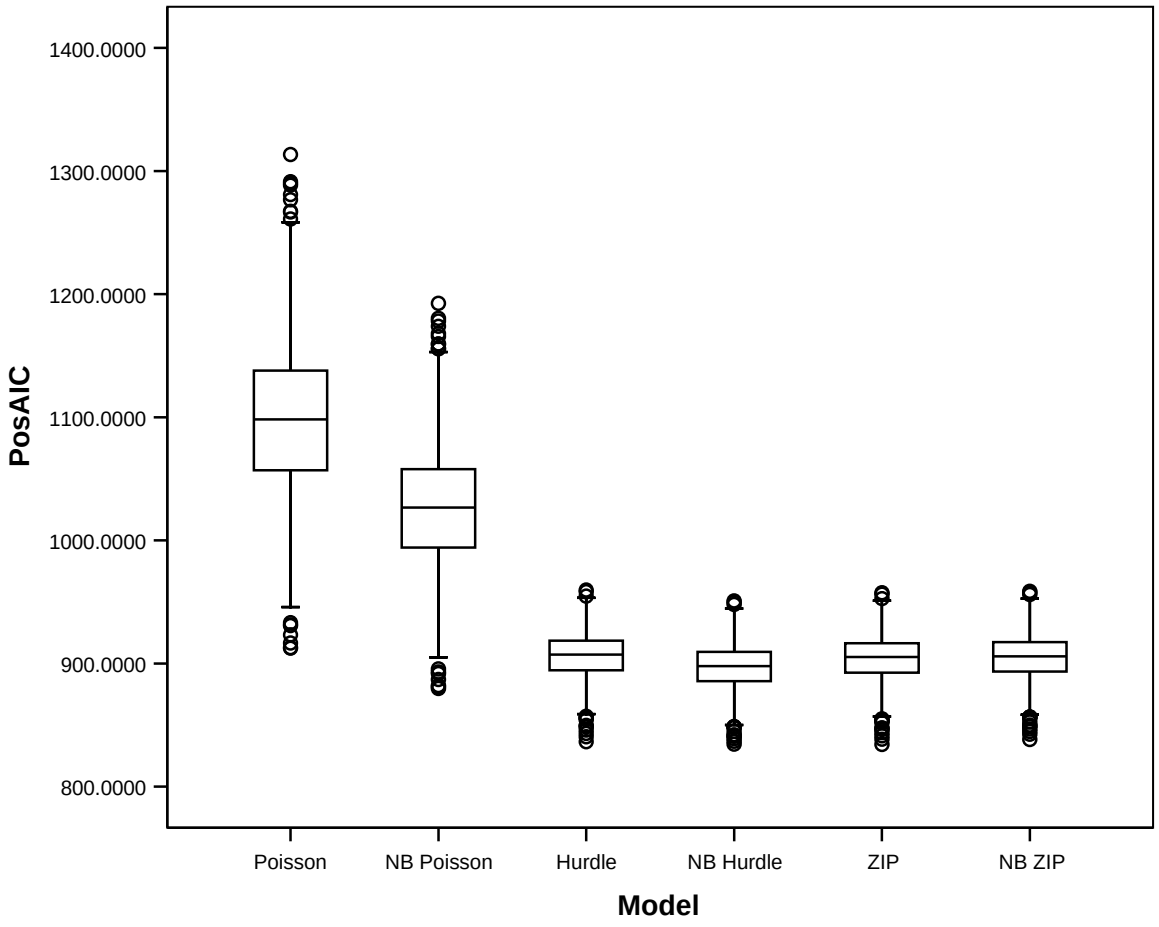

Figure 4-5. Boxplot of AIC's for all models for a .90 proportion of zeros

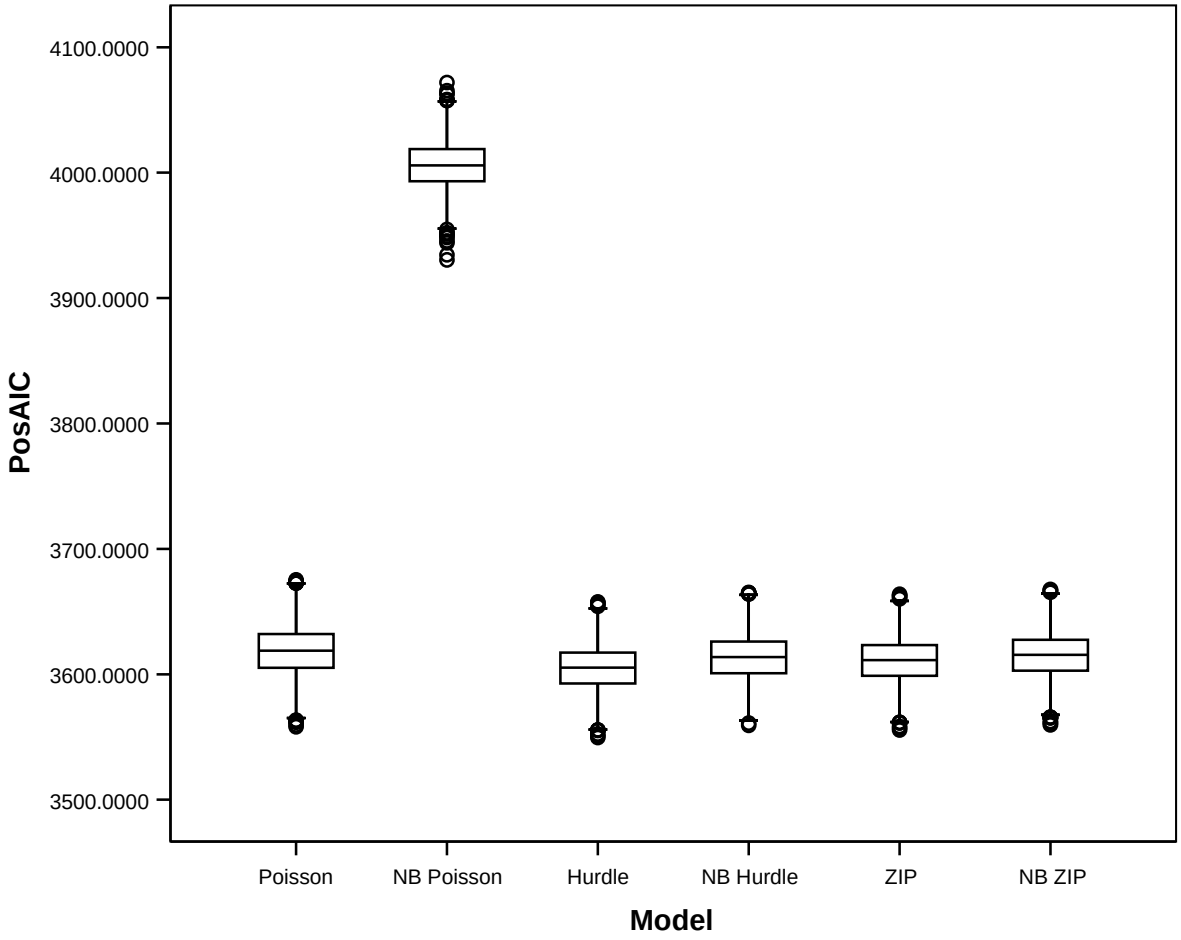

Figure 4-6. Boxplot of AIC's for all models for a .10 proportion of zeros

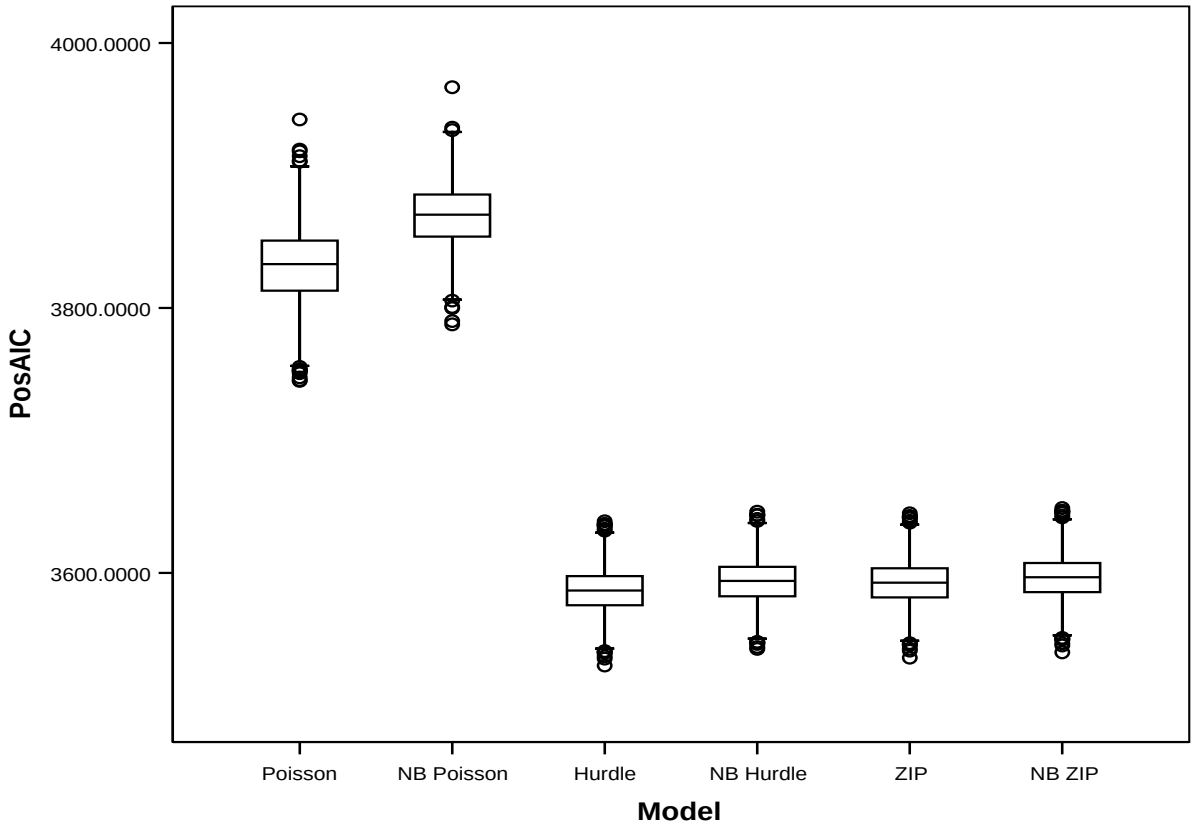

Figure 4-7. Boxplot of AIC's for all models for a .25 proportion of zeros

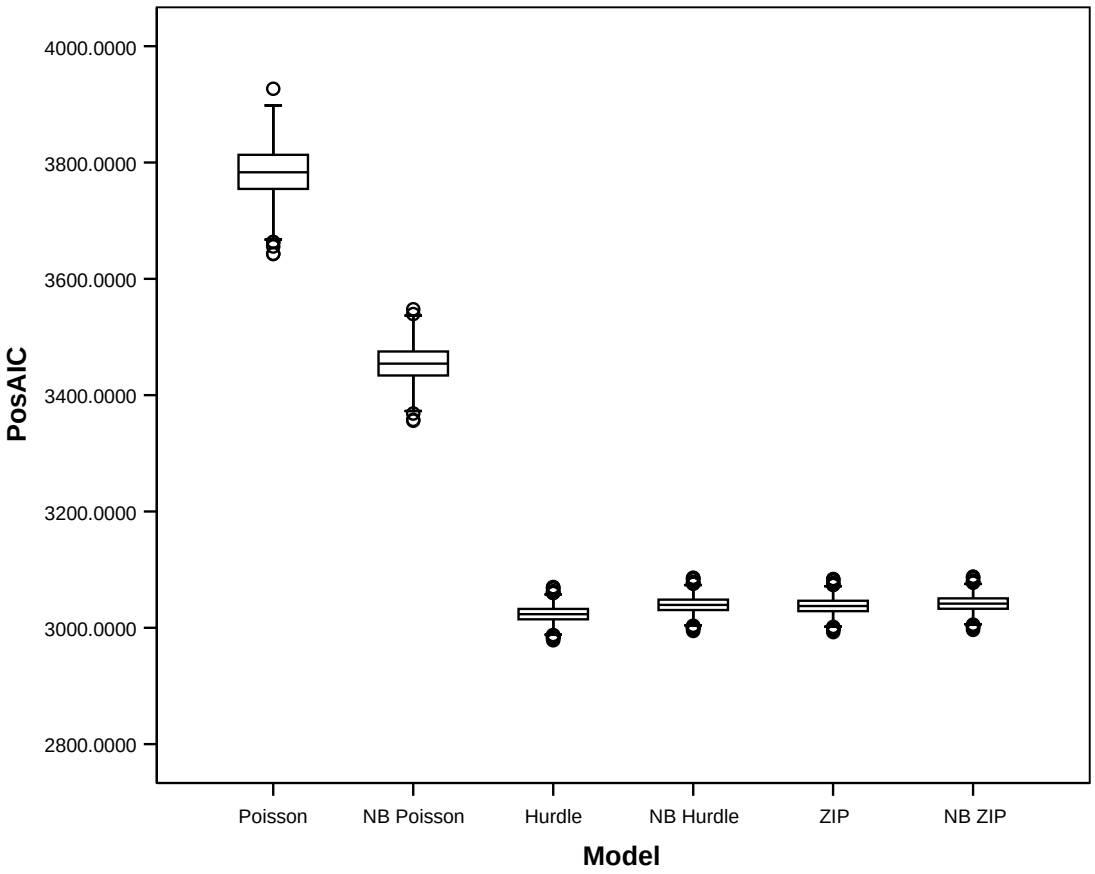

Figure 4-8. Boxplot of AIC's for all models for a .50 proportion of zeros

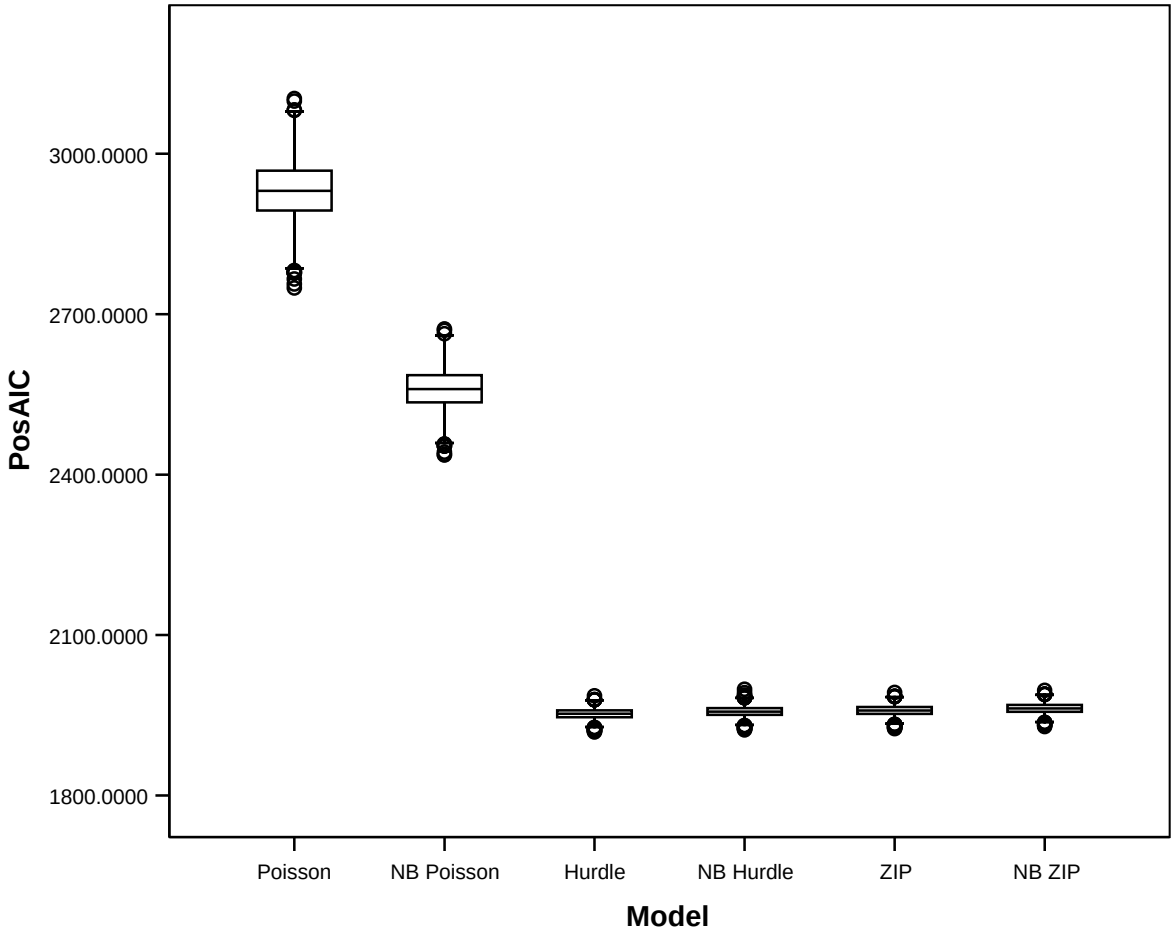

Figure 4-9. Boxplot of AIC's for all models for a .75 proportion of zeros

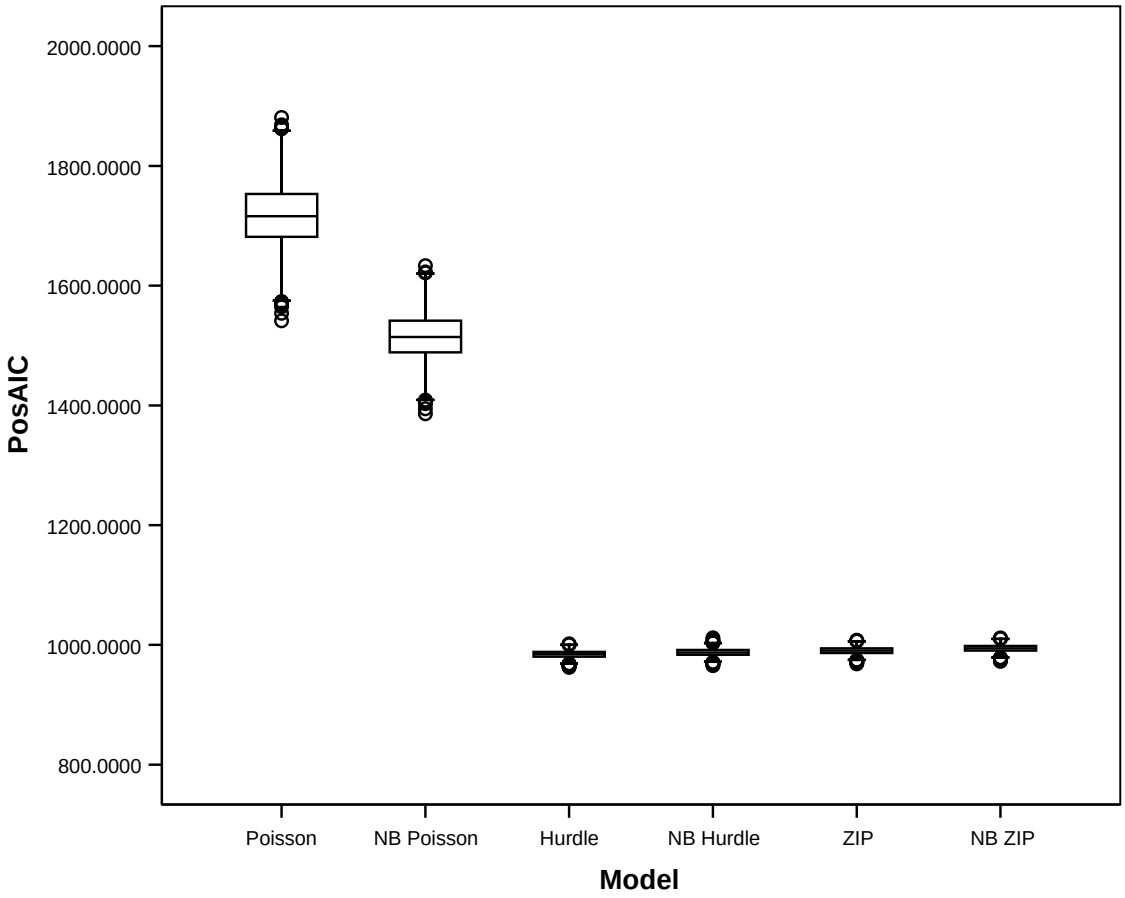

Figure 4-10. Boxplot of AIC's for all models for a .90 proportion of zeros

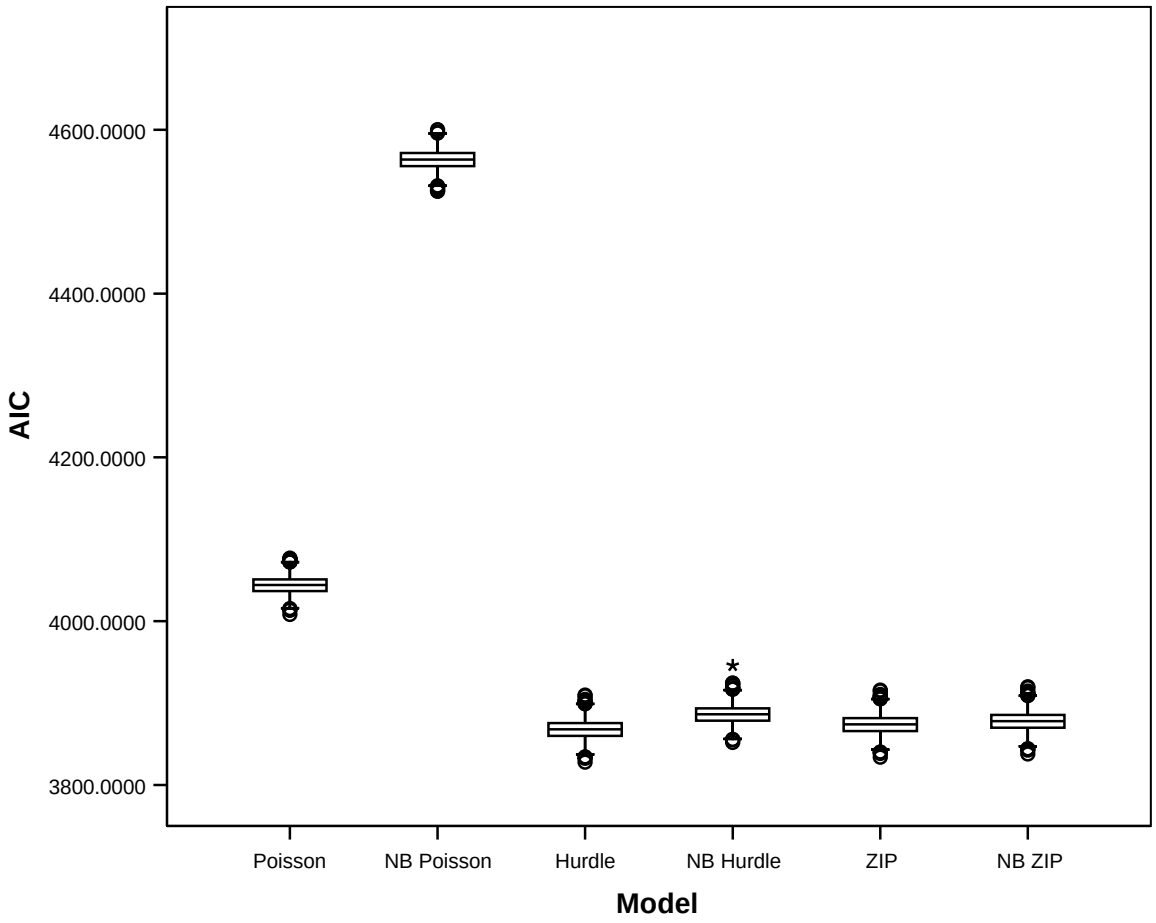

Figure 4-11. Boxplot of AIC's for all models for a .10 proportion of zeros

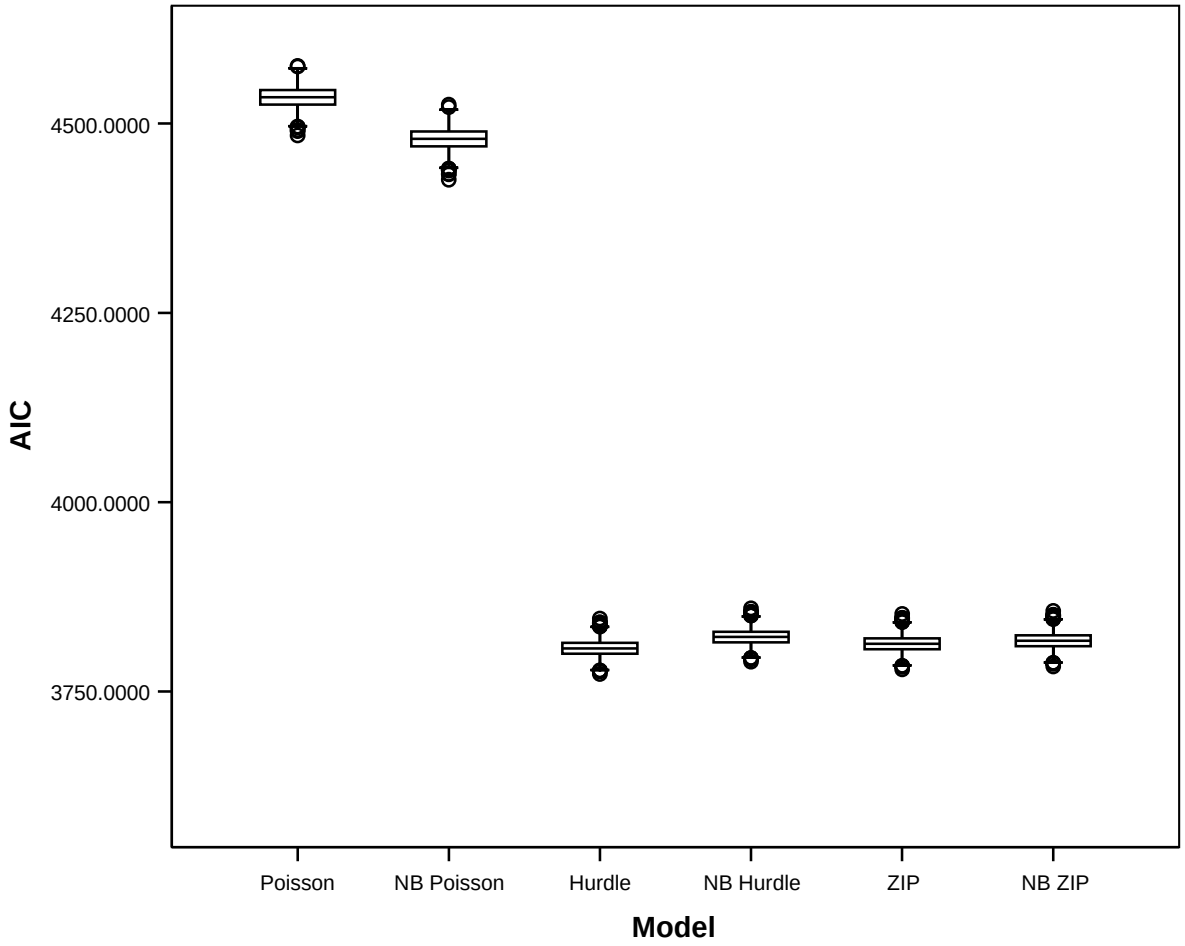

Figure 4-12. Boxplot of AIC's for all models for a .25 proportion of zeros

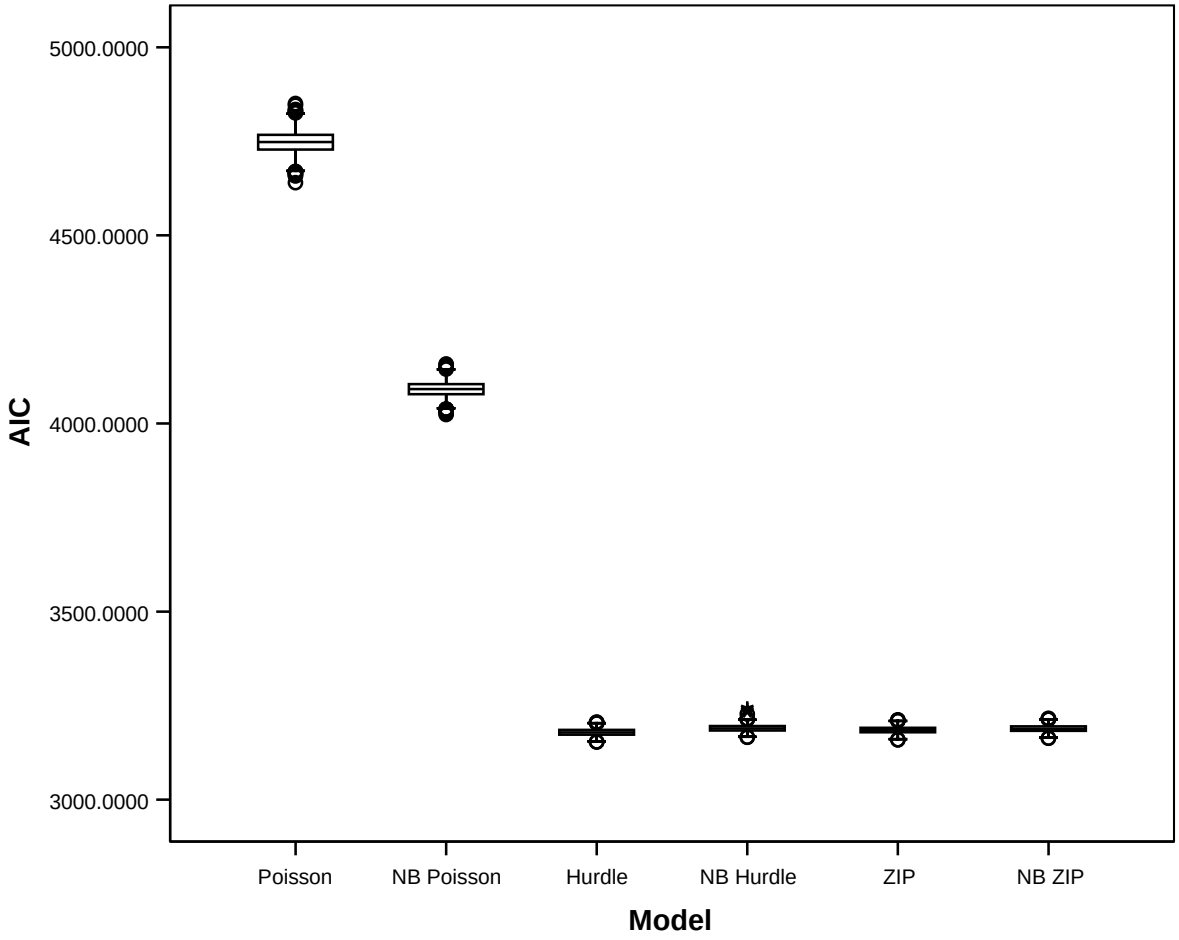

Figure 4-13. Boxplot of AIC's for all models for a .50 proportion of zeros

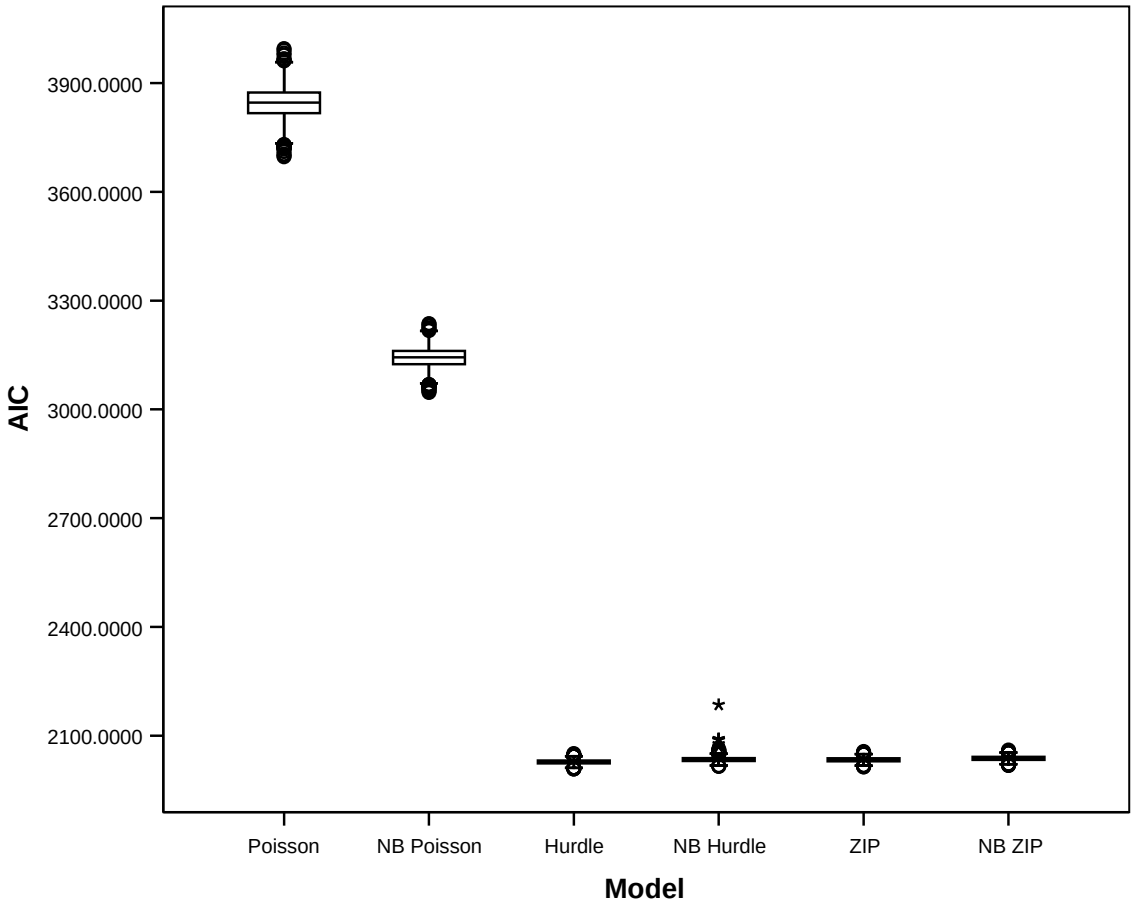

Figure 4-14. Boxplot of AIC's for all models for a .75 proportion of zeros

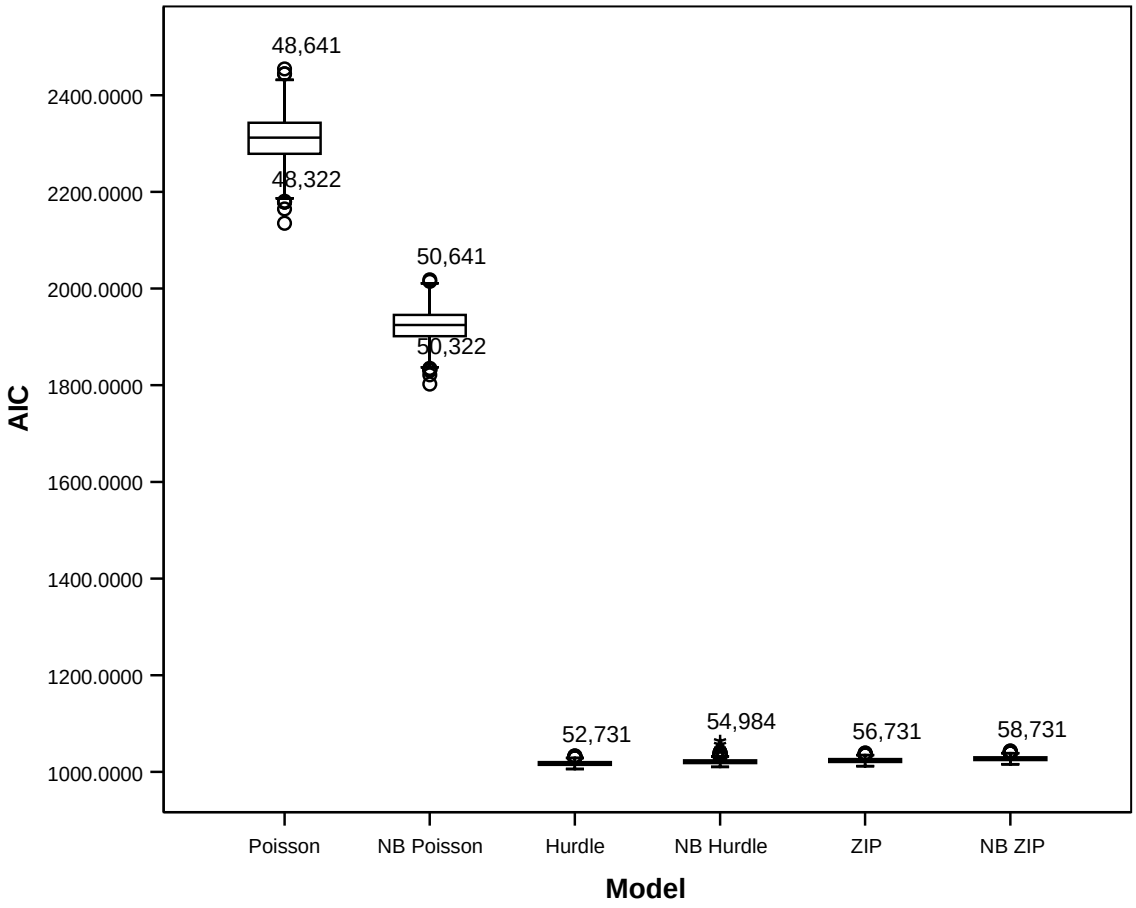

Figure 4-15. Boxplot of AIC's for all models for a .90 proportion of zeros

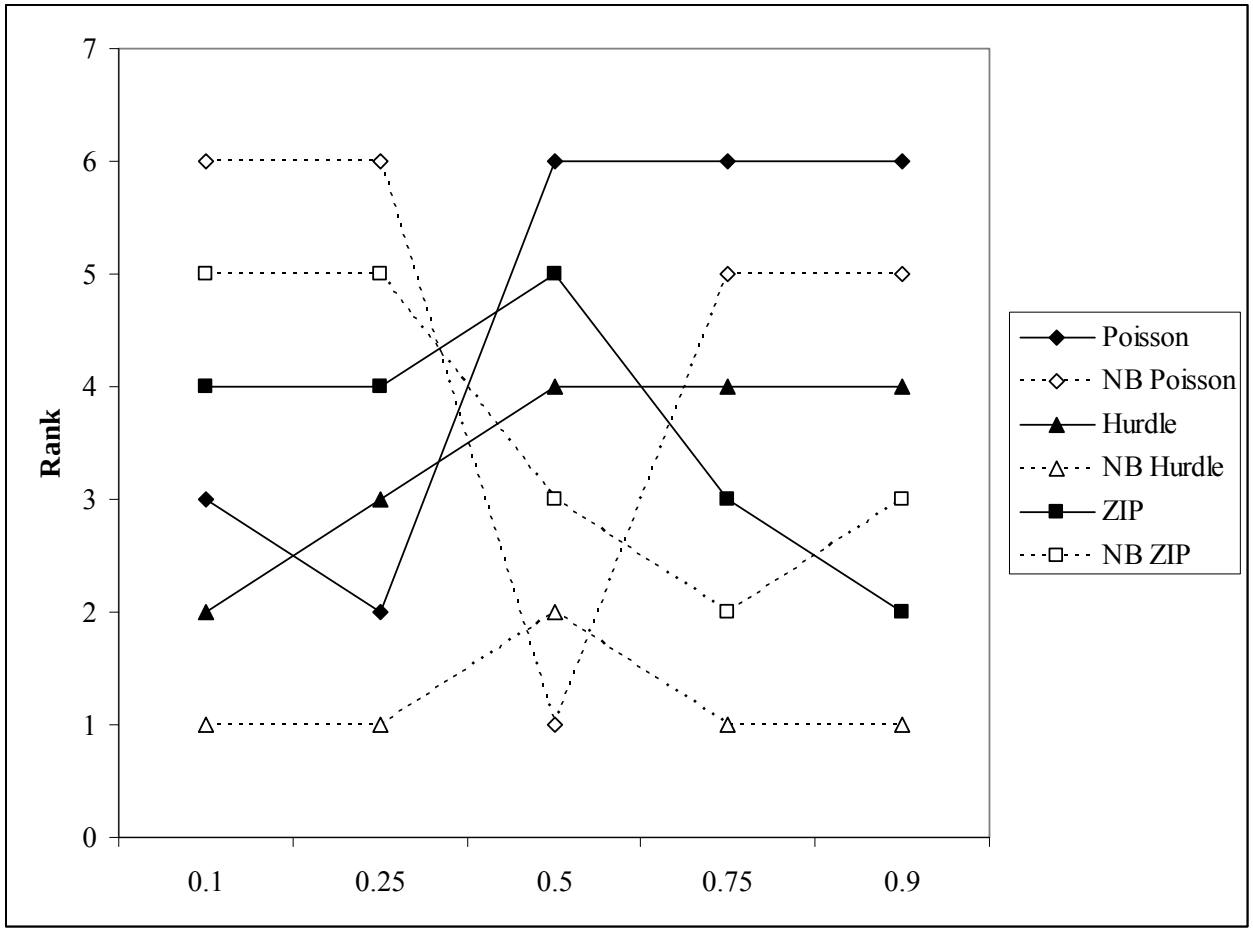

Figure 4-16. AIC rank order for positively skewed distribution models

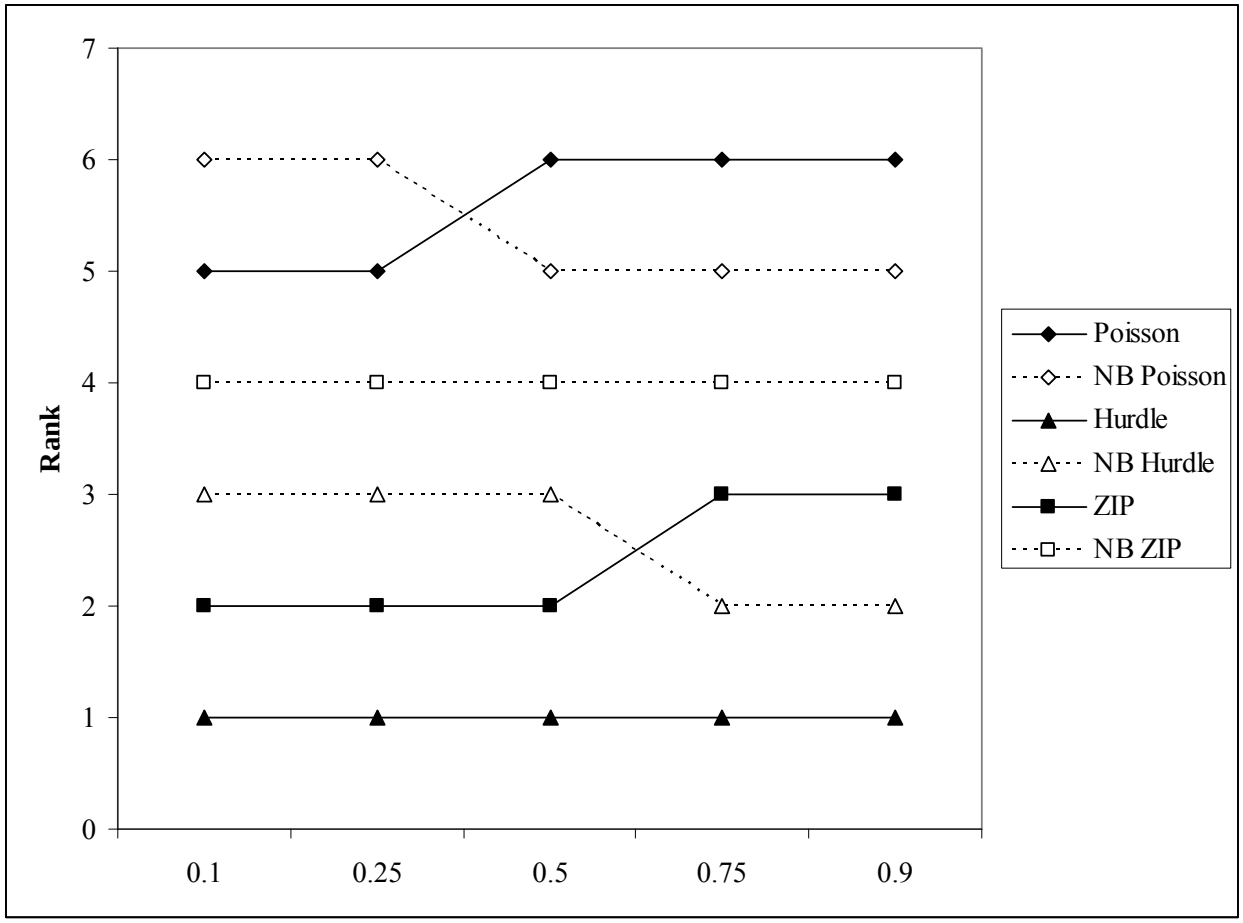

Figure 4-17. AIC rank order for normal distribution models

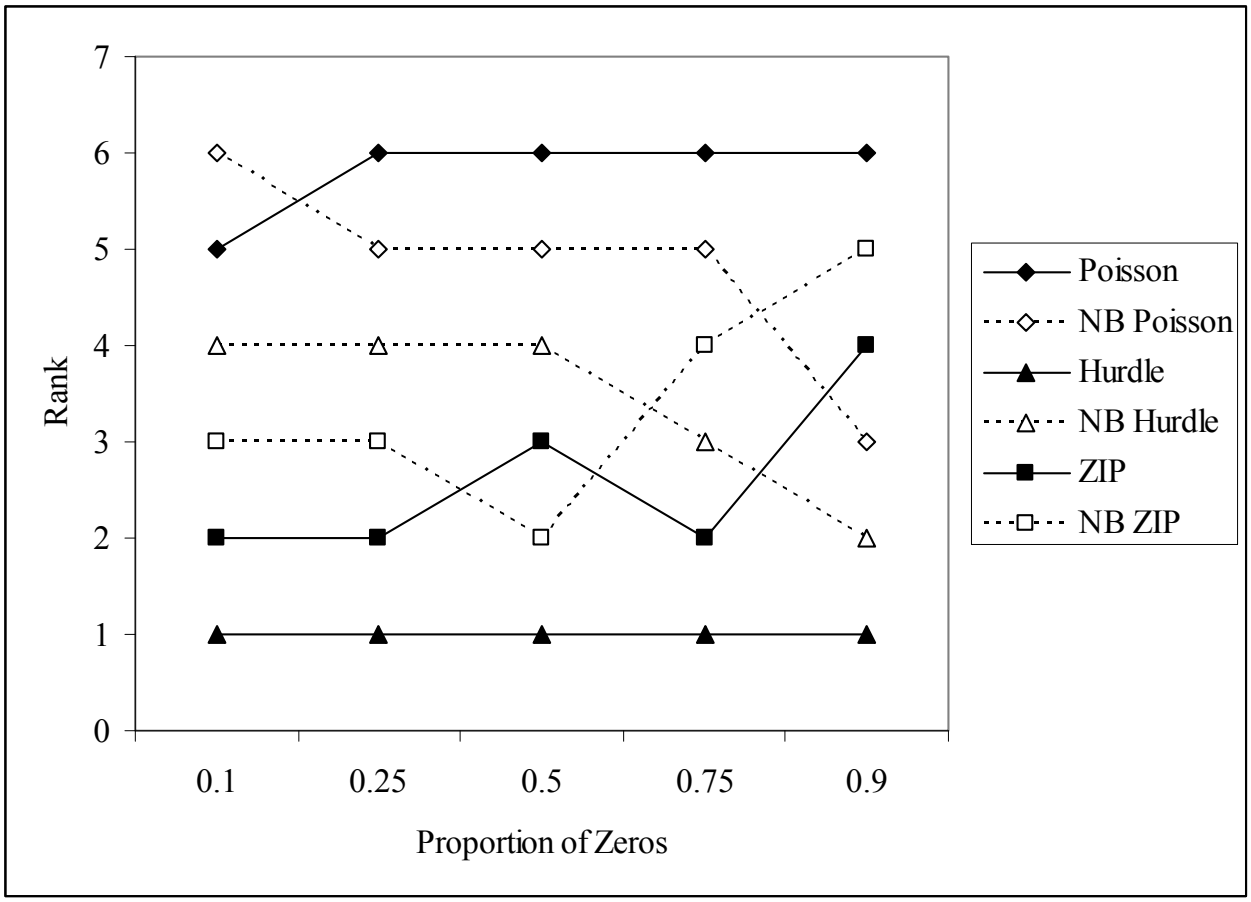

Figure 4-18. AIC rank order for negatively skewed distribution models

CHAPTER 5
DISCUSSION

This research was conducted to determine differences in fit between six models under five conditions of zero-inflation and three conditions of skew for the nonzero distribution.[29] A Monte Carlo study was conducted in which samples of size 1,000 were sampled from distributions with prespecified proportions of zeros and prespecified proportions for the nonzero count levels. Six models were used to analyze the samples with each simulated 2,000 times. Each model was analyzed with each combination of the five proportions of zeros combined with the three levels of skew. The research questions to be answered were as follows:

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated log-likelihood between a) the Negative binomial Poisson model vs. Poisson model; b) the Hurdle model vs. Poisson model?; c) the Negative binomial Hurdle model vs. negative binomial Poisson model?; d) the Negative binomial Hurdle model vs. Hurdle model; and, e) the Negative binomial ZIP model vs. ZIP model?

- Given one two-level categorical covariate with known values and one continuous covariate with known values, what is the difference in the estimated AIC between all models?

The Impact of the Event Stage Distribution

Positively Skewed Event-Stage Distributions

The positively skewed event stage distribution most resembles that which is normally treated as Poisson or negative binomial Poisson. However, departure from these distributions is possible under different conditions of zero-inflation. Further departure may arise depending on the model used to analyze the data. For example, the Hurdle and ZIP formulations treat the event stage as distributed Poisson while treating the transition stage as distributed binomial.

[29] All conclusions from this research are based on these conditions. Hence, conclusions are provisional as there may be other important determinations (e.g., parameter estimates, standard errors, power) not addressed in this research.

As seen in Figure 5-1, the negative binomial Poisson model displayed better fit than the Poisson model only when the proportion of zeros was .50, .75, or .90. This suggests that overdispersion in a positively skewed distribution arises in medium to high proportions of zeros. However, an extremely low proportion of zeros offsets this effect leading to equidispersion and adequate fit for the Poisson model. In other words, the need for a negative binomial model arises as the proportion of zeros increases hence increasing zero-driven overdispersion.

The comparison of the Poisson and Hurdle models presented anomalous findings. At .10 zero-inflation, the Hurdle model displayed superior fit supporting its ability to handle zero-deflation. The same results were found at .50, .75, and .90 zero-inflation supporting the Hurdle models ability to accommodate zero-inflation. However, as seen in Figure 5-1, only 25.2% of the simulations favored the Hurdle model over the Poisson model when the proportion of zeros was .25. It is possible that the proportions for all counts yields a distribution that is distributed Poisson more than a distribution that treats the zeros as distributed binomial.

The comparison of the negative binomial Poisson and negative binomial Hurdle yielded similar findings. However, this time it was at the .50 proportion of zeros that fewer simulations (i.e., 13.3%) favored the more complex model. The reason for this is the addition of the overdispersion parameter for the negative binomial formulations of the model. This parameter accounts for zero-driven overdispersion; however, at .50 zeros, the fit of the overall distribution is reduced by treating the zeros as distributed binomial. The decrease of zero-driven overdispersion at .10 and .25 zeros and the increase of zero-driven overdispersion at .75 and .90 zeros then lend support to the negative binomial Hurdle model. This is illustrated in Figure 5-2.

The negative binomial Hurdle consistently fit better than the Hurdle model until the proportion of zeros was .90. This can be seen in Figure 5-2; at this proportion, the percent of

simulations favoring the negative binomial Hurdle model was 30.7%. For low proportions of zeros, this finding supports Agresti and Min's (2004) assertion that the Hurdle models can accommodate zero-inflation as well as zero-deflation. However, as the proportion of zeros increases, the negative binomial Hurdle would be expected to display better fit due to the increase in zero-driven overdispersion. One explanation is that the zero-driven dispersion at high proportions of zeros is counteracted by the reduction in sample size for the event-stage. For example, if 90 of 100 responses are zero then there are certainly far fewer counts contributing to event-stage model fit.

The fit of the ZIP model formulations, which permit zeros in the event stage distribution, were affected by both the proportions of zeros and the positively skewed distribution. At extremely high or low proportions of zeros (e.g., .10, .90), the majority of the simulations suggest adequate fit for the ZIP model. This is reasonable since, for a .10 proportion of zeros, the probability of a zero in the event stage is low; the results become similar to that of the Poisson model at a .10 proportion of zeros. As was the case when comparing the Hurdle and negative binomial Hurdle models, there are not enough zeros in either of the two parts of the model to necessitate a negative binomial formulation to handle zero- and event-driven overdispersion. For a .90 proportion of zeros, a portion of zeros transferred to the event-stage leads to more skew in the event-stage. In between these conditions, overdispersion arises due to the proportion of zeros justifying the negative binomial formulation. In fact, as seen in Figure 5-3, the negative binomial ZIP displayed significantly better fit than the ZIP model for almost 95% of the simulations when the proportion of zeros was .50.

An alternative explanation to these findings comparing ZIP and negative binomial ZIP model fit pertains to convergence. As displayed in Table 4-42, there were convergence problems

for the ZIP models, especially when the proportion of zeros was .10. Hence, the comparisons are less stable since they are based on a lower number of simulations. This supports Min and Agresti's (2004) findings of superior convergence rates for the Hurdle models over the ZIP models.

The AIC's revealed inconsistent findings compared to those of the deviance tests. This is most likely due to the similarity in AIC's offset by their penalty for the number of parameters in the models. For the nested models, the deviance test provides the more valid interpretation of results. The AIC's can be useful if a) the researcher is unsure of the nature of a dual data generating process or b) whether a model reflecting a dual generating process is necessary. Given this data, the negative binomial Hurdle is superior to the ZIP models. However, the superiority of the other models appears to depend on the proportion of zeros. At a low proportion of zeros, the Hurdle model was superior to the ZIP models; although, at a high proportion of zeros, the ZIP model was superior to the Hurdle models.

There were several situations in which the Poisson models were superior to the two-part models suggesting that the two-part models may not be necessary in some conditions. At a very low proportion of zeros (i.e., .10), the Poisson model was superior to both ZIP models. At the .25 proportion of zeros, the Poisson model was superior to both the ZIP models and the Hurdle model. The negative binomial Poisson model is typically a poor choice over the two-part models and the Poisson model with one major exception. At a proportion of zeros of .50, the negative binomial Poisson model can be considered the best fitting model based on the penalty for model complexity provided by the AIC's.

The results are in line with the findings of Lambert (1992). At a .718 proportion of zeros, her results revealed that the ZIP model was superior to the negative binomial Poisson model,

which was superior to the Poisson model. The present study found these results at a .75 proportion of zeros. Overall, the results imply that zero-inflation and zero-deflation are best accommodated for by using a negative binomial Hurdle model; however, when half of the data is zeros, the negative binomial Poisson model will suffice.

Greene (1994), using a heavily positively skewed distribution with a .894 proportion of zeros found similar results for preferring the negative binomial ZIP model over the ZIP model. However, for the present study, at a proportion of zeros of .90, this was only true for .133 of the simulations; this may be due to the convergence problems.

Mullahy (1986) found adequate fit for the Hurdle model only at a .26 proportion of zeros. At proportions of .41 and .62, the fit was no longer adequate. This is in agreement with the findings of the present research. At low proportions of zeros with a positively skewed distribution and in terms of AIC's, the Hurdle model displayed fit only improved by the negative binomial Hurdle model. However, the fit for the Hurdle model became worse as the proportion of zeros increased.

Overall, these results provide suggestions for analyzing data with various proportions of zeros and a positively skewed event-stage distribution. First, if there is no reason to justify a two-part model then it is not necessarily prudent to assume a negative binomial Poisson distribution to accommodate zero-driven overdispersion. The results of this study suggest that the negative binomial Poisson model is superior to the Poisson only when the proportion of zeros is greater than or equal to .50. Second, if justifying separate processes generating the zeros and nonzeros then both the Hurdle and negative binomial Hurdle should be analyzed and compared at proportions of zeros around .90 since here there appears to be a drop in the need for the negative binomial Hurdle model. Third, it appears that the need for the negative binomial ZIP model

declines as the proportion of zeros increases from .50 to .90. However, all interpretations should be treated with caution due to the severe convergence problems at the .10 and .25 proportion of zeros conditions. Hence, if the researcher acknowledges a dual data generating process but is uncertain about the source, the ZIP models are more valid since they don't exclude zeros from the event-stage that may be present. However, there is a greater probability that the analysis will not converge. This again supports Min and Agresti's (2004) assertion of Hurdle model superiority in the condition of zero-deflation. Finally, under certain conditions, two-part models aren't necessary. In particular, a Poisson model may fit as well as a Hurdle model at .25 zeros, and a negative binomial Poisson model may fit as well as a negative binomial Hurdle model at .50 zeros.

Normal Event-Stage Distributions

Regardless of the proportion of zeros, the Hurdle model fits best when the event stage distribution is normal. The negative binomial Hurdle is not necessary. However, the results should be interpreted cautiously approximately since, as displayed in Table 4-43, 25% of simulations failed to converge in the .25 and higher proportion of zeros conditions. This is an interesting finding since previous research (Min & Agresti, 2004) and these findings displayed convergence problems for the positively skewed distribution.

If opting for a two-part model in which zeros are permitted in the event stage, the superior model is the ZIP model. Hence, the statistical analyses suggest that, regardless of the proportion of zeros, negative binomial formulations are not necessary for both the ZIP and Hurdle models.

If a two-part model is not necessary then the model depends on the proportion of zeros. As was the case with a positively skewed distribution, negative binomial Poisson models are not necessary when the proportion of zeros is .10 or .25. It is at proportions of .50, .75, and .90 zeros that the negative binomial Poisson model fits the data better than the Poisson model.

It is important to note that model superiority does not inform researchers as to the extent of that superiority. The descriptive AIC values suggest better fit for the Hurdle over the other two-part models; however, the AIC's for the other two-part models do not differ by much. This is most prevalent at the .10 proportion of zeros condition; the relative advantage of the Hurdle model over the other two-part models declines as the proportion of zeros increases. At the .90 proportion of zeros condition there is barely any distinction between the four two-part models at all.

The discrepancy between Poisson model and negative binomial Poisson models is consistently large across proportions of zeros. The only condition where they approach the AIC's of the two-part models is at the .10 proportion of zeros condition. Overall interpretations can then be made. First, the Hurdle model is the best-fitting model of all two-part models; however, the extent of this fit is greatest at low proportions of zeros. Second, negative binomial Poisson models are advantageous at medium to high proportions of zeros. Finally, when unsure about whether to use a two-part model, a Poisson model will suffice at low proportions of zeros.

Slymen, Ayala, Arredondo, and Elder (2006) compared the Poisson model to the negative binomial Poisson model and the ZIP model to the negative binomial ZIP model. The proportion of zeros was .824, and the event stage distribution was uniform. The present study, at a .90 proportion of zeros, has an event stage distribution that appears uniform given the reduction in event stage frequencies from a possible 1,000 to 100. The results between the studies were similar in that the negative binomial model is superior to the Poisson model. Also, similar to their findings, there was no drastic distinction between the ZIP and negative binomial ZIP models at this proportion of zeros.

Böhning, Dietz, Schlattmann, Mendonça, and Kirchner (1999) compared the Poisson and ZIP models under two proportions of zeros both close to .25 and with normally distributed event stages. There results are similar to those of the present study. The ZIP model is superior to negative binomial ZIP. However, this isn't surprising since the ZIP mode is consistently superior to the negative binomial ZIP model regardless of the proportion of zeros. Only when adjusting for model complexity via AIC's do the model become more similar in fit. Table 4-38 summarizes these findings.

Negatively Skewed Event-Stage Distributions

No previous research was found in which the event stage distribution was negatively skewed. It is possible that researchers begin with the notion of zero-inflation, which, due to the count nature of the data, leads to considering the Poisson distribution. Introductory texts introduce a Poisson distribution with a low mean yielding a positively skewed distribution. Hence, negatively skewed zero-bounded distributions receive little consideration.

Distributions displaying both zero-inflation and a negatively skewed event stage are certainly feasible. For example, the response variable may be student's number of days absent from an honors course. A proportion of these students should have no absences. Beyond zero, there may be a relationship between enrolling in the honors course and participating in extracurricular activities requiring absences. These students would then have more absences than the others leading to a negatively skewed distribution.

The negative binomial Poisson model displayed significantly better model fit than the Poisson mode for all conditions except when the proportion of zeros was .10. The Hurdle model fit better than the Poisson model, and the negative binomial Hurdle model fit better than the negative binomial Poisson model for all simulations. The negative binomial Hurdle model fit better than the Hurdle model for all simulations except when the proportion of zeros was .75. In

this condition, the negative binomial Hurdle model fit better than the Hurdle model for 90.7% of the simulations. As was found for the normal distribution conditions, conclusions regarding the negative binomial Hurdle model should be treated cautiously as convergence failures were approximately 25% over all proportion of zeros conditions. Finally, the negative binomial ZIP model never displayed significantly better fit than the ZIP model.

These findings are interesting in that, although the proportion of ones is approximately four times the proportion of zeros, the overall distribution (including the zeros) is negatively skewed. In other words, the serious zero-deflation may be disguised as part of one negatively skewed distribution. Hence, the implications of these results apply to zero-bounded data that appear to be the inverse of the typical Poisson distribution. The negative binomial Hurdle model could be used to analyze such a distribution provided that the zeros can be viewed as arising from a completely different mechanism than the nonzeros. However, the simpler ZIP model could be used if there is reason to consider two zero generating processes. Both of these models provide better fit over the Poisson.

No previous research was found to address a negatively skewed event stage distribution. However, there did appear to be one incongruent finding. Several articles reported that the Hurdle was advantageous over the ZIP model for zero-deflation since, while both give nearly equivalent results, the ZIP model often has problems with convergence (Min & Agresti, 2004; Min, 2004). For the negatively skewed distribution as well as for the normal distribution, the negative binomial ZIP models failed to converge only once out of a total of 20,000 simulations while the Hurdle models failed to converge 3,899 times. This is seen in Table 4-44. It is possible that the convergence advantage for the Hurdle models only applies to a positively skewed event-stage distribution. However, the present research only found convergence problems for the

Hurdle model at low proportions of zeros with 112 successful convergences out of 2,000 for the .10 proportion of zeros and 1,850 successful convergences out of 2,000 for the .25 proportion of zeros.

Summary of Findings

One of the most interesting findings pertained to the Poisson and negative binomial Poisson models. When the event-stage distribution is positively skewed or normal, the negative binomial Poisson model is not necessary until the proportion of zeros is .50. For the negatively skewed distribution, the negative binomial Poisson is not necessary when the proportion of zeros is .10. The results are not surprising or the positively skewed and normal distributions; the proportion of zeros is so low that the event-stage distribution not overdispersed. For the negatively distribution, the interpretation is suspect since, ignoring the zeros, the event-stage distribution has no resemblance to the Poisson distribution. However, at least for the positively skewed and normal distributions, these results present some guidelines for selecting either the Poisson or more complicated negative binomial Poisson model.

Another interesting finding pertains to the Hurdle and negative binomial Hurdle models. For the positively skewed distribution, the negative binomial Hurdle model should be chosen regardless of the proportion of zeros. This was also true for the negatively skewed distribution. However, for the normal distribution, the more complicated negative binomial Hurdle model is not necessary. This provides a guideline for choice between the Hurdle and negative binomial Hurdle models for the distributions if willing to trust the statistical conclusion validity of 75% successful convergences for the normal and positively skewed distribution comparisons.

Finally, this research showed that the negative binomial ZIP model is never necessary when the event-stage distribution is normal or negatively skewed. In fact, even for the positively skewed distribution, it may not be necessary except at .10, .25, and .90 proportions of zeros. The

negative binomial ZIP model appears to be superior when the event-stage distribution is positive and when there is moderate to moderately-high zero-inflation but not extreme zero-inflation.

Limitations

Discrete Conditions

The results generated by this Monte Carlo simulation derive from prespecified discrete conditions. There were five prespecified proportions of zeros and three prespecified levels of skew. The former was fixed for each dataset, and the latter was determined by randomly selecting from prespecified pseudo-population probabilities for each of the five count levels. The covariates were fixed as one continuous variable distributed standard normal and one categorical variable with two levels each at $n = 500$ and alternating consecutively in the dataset.

The interpretation of results is limited to these discrete conditions. For example, for the negatively skewed event count distribution, the Poisson model fit better than the negative binomial Poisson model at a .10 proportion of zeros. However, the reverse was true at a .25 proportion of zeros. This is not to say that the Poisson model has better fit until the proportion of zeros is .25. It could be the case that the negative binomial Poisson has better fit at a proportion of .11 zeros. Further, the results are generalizable to the extent that applied research maintains the same proportions for event stage counts (i.e., skew) and covariate characteristics as used in this study.

Convergence and Optimization

Optimization is a mathematical process of finding the best solution to a function. Statistically, optimization is achieved through an iterative process until convergence is reached resulting in the optimum values. This topic was previously discussed in the context of maximum likelihood estimation.

Different optimization procedures can produce different results and different rates of convergence. The Hurdle model requires complex optimization since there are two models to be optimized. The ZIP model is even more complex since the two models are not independent of one another. The negative binomial ZIP model is most complex since the event stage distribution requires the addition of a dispersion parameter.

There is no perfect optimization procedure that finds the best solution within the most reasonable amount of time for all sets of data (SAS Institute, 2000). For the SAS programming language, these zero-inflated models would typically be analyzed using PROC NLMIXED. The default optimization procedure here is the Quasi-Newton method which does not require computation of second-order derivatives. Hence, it has the advantage of finding solutions rather quickly for moderately large problems. However, it does not consider the boundary constraints present in the zero-inflated data. In fact, the Quasi-Newton method is not even included as an option in the R procedures for the Hurdle and ZIP models. Here, the default is the Nelder-Mead Simplex Optimization. This method does not require any derivatives and permits boundary constraints. Its process increases speed of convergence while large sample sizes (e.g., $N > 40$) increase convergence time and decrease the probability of generating a precise result (SAS Institute, 2000). This may then explain some of the discrepancy in findings between this study and that of Min and Agresti (2004) and Min (2004) whose studies used the NLMIXED procedure with its default Quasi-Newton optimization method.

Underdispersion

Strictly speaking, conditions such as a .10 proportion of zeros with a negatively skewed event stage distribution can lead to a variance that is less than its mean. As such, it is expected that the models outperform their negative binomial counterparts. However, a dispersion

parameter that accounts for underdispersion may reveal a model that has superior fit to these models. Further, the results are likely to vary between the Poisson, Hurdle, and ZIP formulations.

Interpreting Model Fit

This research used the deviance statistic and Akaike's Information Criterion to examine fit between different models. There are certainly other alternatives. Vuong (1989) proposed an inferential likelihood ratio test for comparing nonnested models. Alternatives to the AIC include the Bayesian Information Criterion, which is typically used for unbalanced sample sizes and the Schwarz Criterion, which is typically used for hierarchical models.

Other models

ZIP and Hurdle models have become the most common models for analyzing zero-inflated data. However, as previously discussed, there are many other models to choose. Some of these include Lambert's (1989) ZIP(τ) to accommodate correlation between the zeros and nonzeros, other links functions for the event-stage (e.g., log-gamma, log-log), and Ridout, Demétrio, & Hinde's (1998) method for avoiding some of the distributional assumptions.

Another potential model to analyze zero-inflated count data is the latent class model. The model is similar to a general linear model in that one may posit group membership. However, in latent class modeling, the groups are unobserved. Hence, it is able to provide estimates for potential classes that may explain heterogeneity in responses (Ding, 2006). It may be very useful for researching zero-inflated count data due to the two potential sources of zeros. In fact, d'Uva (2006) proposed a Hurdle variation for latent class modeling in health care utilization research.

Validity of Model-Fitting and Model-Comparisons

The purpose of comparing models is to determine which has the best fit. This is done by selecting the model that minimizes the loss of information. As previously mentioned, this is often done by comparing statistics such as the log-likelihood, deviance, and Akaike's Information

Criterion (AIC). However, any model will fit perfectly or nearly perfectly given enough parameters (Lord, Washington, & Ivan, 2006). Further, there is always the possibility that a different set or combination of parameters will yield the same fit statistics.

When comparing the Poisson and zero-inflation models, it is important not to lose sight of validity concerns. Ultimately, the best model will not be the one with the best fit. Rather, it will be the one that leads to correct inferences, interpretations, and decisions.

Although one may not know the model specification that will lead to enhanced statistical conclusion validity, it is possible to maintain several inferential principles as clarified by Burnham and Anderson (2001) and reviewed by Mazerolle (2004). First, the ideal model will be simple and parsimonious in line with Ockham's Razor, which states that, all else being equal, the simplest model is usually the correct model. The use of Akaike's Information Criterion as a method for adjusting for model complexity is one example of striving for parsimony (Lord, Washington, & Ivan, 2006). Another example is favoring a simpler model over a more complex model when the fit is nearly identical: "Our choice of the ZIP model over the [negative binomial ZIP] model was based on parsimony since they provide similar fit. The ZIP model does not include the random error term that allows the conditional variance of y to exceed the conditional mean. But the interpretations of the model regression parameters are the same for both models" (Slymen, Ayala, Arredondo, & Elder, 2006, p.8).

Second, the researcher should maintain several hypotheses noting that there is no perfectly true model. Third, since validity is a process of accumulating evidence, conclusions should be substantiated by continuing efforts to confirm the strength of evidence using various techniques such as replication and cross-validation.

Suggestions for Future Research

The results of this study provide ample suggestions for future research. As previously mentioned as limitations, it is suggested that future research consider other proportions of zeros and event stage distributions, underdispersion adjustments, different optimization procedures, and other models. This research should also be extended to ordinal data by developing a model that combines the ZIP and Hurdle models with a cumulative model (e.g., proportional odds model). The research should also be extended to latent outcomes possibly by specifying a probit distribution. Although recent research has been extended to incorporate random effects (e.g., Min & Agresti, 2005; Hall & Zhang, 2004; Hall, 2004; Hall, 2002; Olsen, 1999), these findings should be examined in the context of varying proportions of zeros and event stage distributions

Application in Educational Research

Models for zero-inflation have been applied in a wide variety of fields from the health sciences to economics to transportation. However, no research was found to apply models for zero-inflation to the field of education. There are plenty of opportunities to analyze data with such methods (e.g., retention, substance abuse, disciplinary action).

In an era of high-stakes accountability, it is imperative that educational researchers use the most sophisticated methods available. This is especially important since the results of these studies lead to decisions directly affecting human lives. It would be fallacious to treat zero-inflated count data as purely continuous analyzed via ordinary least squares regression. A higher level of sophistication would lead to specification of a Poisson distribution with a log link. However, this would result in responses of zero included with the other responses leading to incorrect inferences due to structural and/or random reasons for the responses of zero.

It would be more judicious to first consider the nature of the zeros and then consider the use of the two-part models. Unfortunately, the application of sophisticated methods tends to

become more pervasive as the methods become simple to apply. As such, educational research programs should promote the existence of these methods, and software programs should provide prepackaged functions for ease of application.

Major Contribution of Findings

The most important finding of this dissertation was that the results of the simulation were in agreement with previous findings when the proportion of zeros and event stage distribution were similar. Hence, anomalous findings between previous research can be attributed to differences in proportions and differences in the event stage distribution. The intention is that researchers become more aware of how different proportions and distributions affect the findings when comparing models such as the Poisson, ZIP, and Hurdle models. For example, as discussed, the negative binomial Poisson model can be superior to the two-part models (e.g., positive event stage distribution with a .50 proportion of zeros), and the relative fit of the ZIP model and negative binomial ZIP model can depend strongly on the proportion of zeros (e.g., for a negative event stage distribution).

Ultimately, the choice should be guided by the data generating process. Data with purely structural zeros should not permit zeros in the event stage, data with structural and random zeros should not have those zeros completely separate from the counts. The findings of this dissertation help guide researchers when uncertain or when questioning the need for a particular model

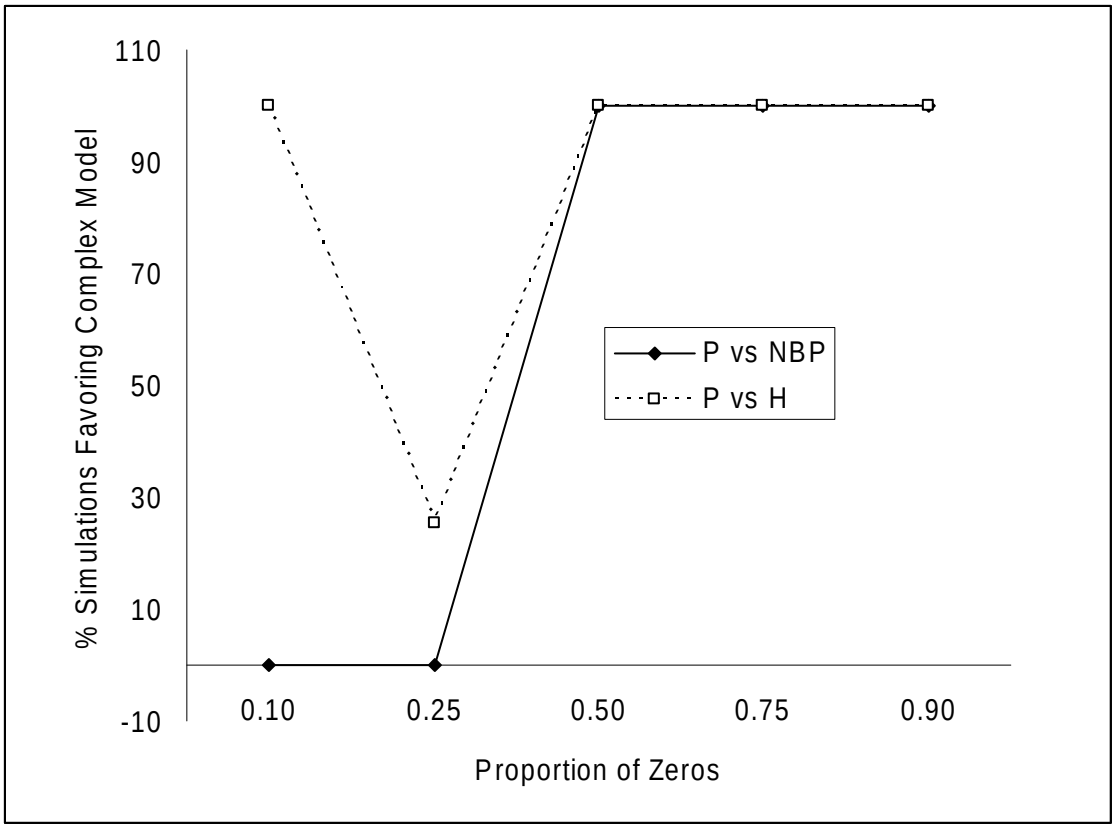

Figure 5-1. Poisson, NB Poisson, and Hurdle over all proportions of zeros

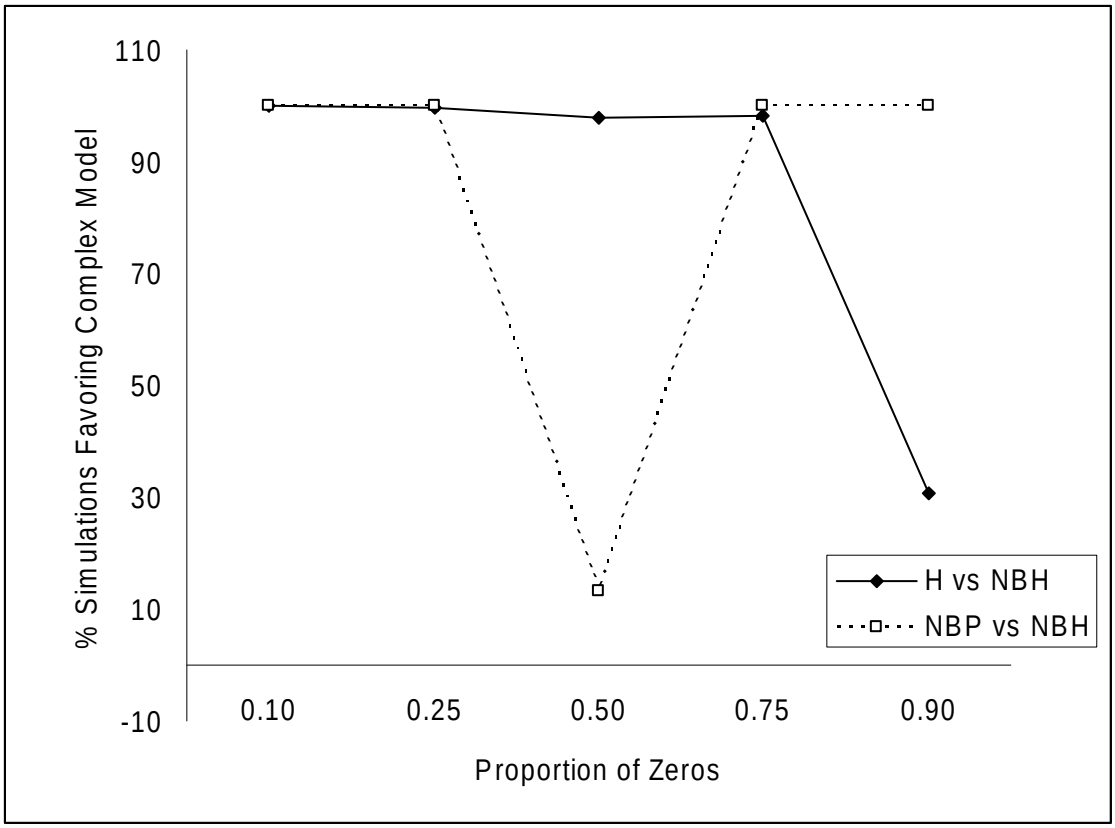

Figure 5-2. Hurdle, NB Hurdle, and NB Poisson over all proportions of zeros

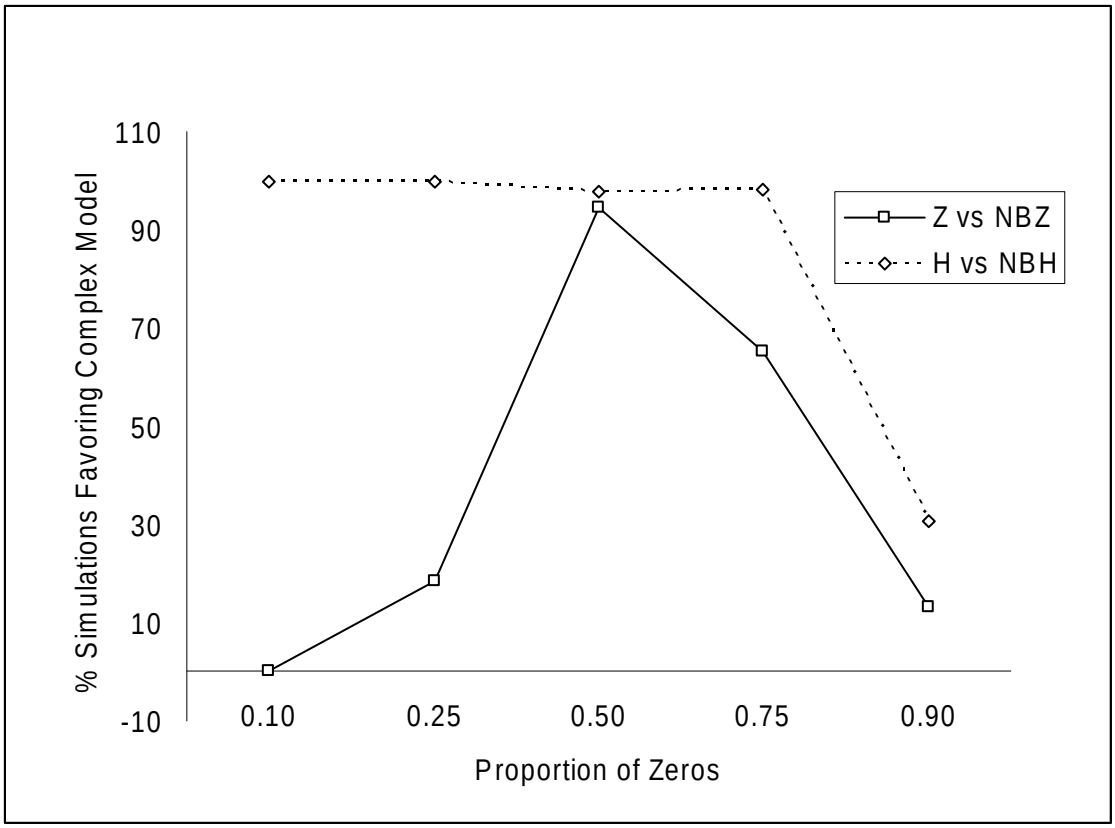

Figure 5-3. ZIP, NB ZIP, Hurdle, and NB Hurdle over all proportions of zeros

LIST OF REFERENCES

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley and Sons.

Akaike, H. (1972). Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*. B. N. Petrov, & F. Csaki (eds.). Akademiai Kiado: Budapest.

Atkinson, K. A. (1988). *An Introduction to Numerical Analysis* (2nd ed.). New York: John Wiley and Sons.

Arulampalam, W. & Booth, A. (1997). Who gets over the training hurdle? A study of the training experiences of young men and women in Britain. *Journal of Population Econometrics, 10,* 197-217.

Berk, K. N., & Lachenbruch, P. A. (2002). Repeated measures with zeroes. *Statistical Methods in Medical Research, 11,* 303-316.

Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L, & Kirchner, P. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Association, Series A, 162,* 195-209.

Bonate, P. L. (2001). A brief introduction to Monte Carlo simulation. *Clinical Pharmacokinetics, 40,* 15-22.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs* Chicago, IL: Scientific Software International, Inc.

Burnham, K. P., and Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research, 28,* 111-119.

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

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

Clarke, K. A. (2001). Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science, 45,* 724-744.

Cohen, A. (1954). Estimation of the Poisson parameter for truncated samples and from censored samples. *Journal of the American Statistical Association, 49,* 158-168.

Cragg, T. C. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica, 39,* 829-844.

Crepon, B., & Duguet, E. (1997) Research and development, competition and innovation: Pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data with heterogeneity, *Journal of Econometrics, 79,* 355-378.

Delucchi, K. L., & Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: Working with many zero values. *American Journal of Psychiatry, 161,* 1159-1168.

D'Uva, T. B. (2006). Latent class models for utilisation of health care. *Health Economics, 15,* 329-342.

Desouhant, E., Debouzie, D., & Menu, F. (1998). Oviposition pattern of phytophagous insects: On the importance of host population heterogeneity. *Oecologia, 114,* 382-388.

Ding, C. S. (2006). Using regression mixture analysis in educational research. *Practical Assessment, Research & Evaluation, 11*, p.1-11.

Dobbie, M. J., and Welsh, A. H. (2001). Modeling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics, 43,* 431-444.

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. *A comparison of alternative models for the demand of medical care.* RAND corporation: Santa Monica, CA.

El-Shaarawi, A. H. (1985) Some goodness-of-fit methods for the Poisson plus added zeros distribution, *Applied and Environmental Microbiology, 5,* 1304-1306.

Famoye, F., & Singh, K. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data, *Journal of Data Science, 4,* 117-130.

Greene, W. H. (1994). *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.* (New York University Department of Economics Working Paper 94-10). Retrieved May 16, 2006, from http://ideas.repec.org/s/ste/nystbu.html

Gupta, P. L., Gupta, R. C., & Tripathi, R. C. (1996). Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis, 23,* 207-218.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. *Journal of Applied Econometrics, 12,* 225-242.

Hagedorn, M., Montaquila, J., Vaden-Kiernan, N., Kim, K., & Chapman, C. (2004). *National Household Surveys of 2003: Data File User's Manual, Volume I.* (NCES 2004-101). Washington, DC: U.S. Department of Education. National Center for Education Statistics.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics, 56,* 1030-1039.

Hall, D. B., & Berenhaut, K. S. (2002). Score test for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. *The Canadian Journal of Statistics, 30,* 1-16.

Hall, D. B., & Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modelling, 4,* 161-180.

Heilborn, D. C. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data. Unpublished technical report, University of California, San Francisco, Dept. of Epidemiology and Biostatistics.

Insightful. (2005). S-Plus (Version 6.0). Seattle, WA: Insightful.

Jackman, S. (2006, November). The pscl package. Retrieved May, 2006 from http://cran.r-project.org/doc/packages/pscl.pdf

Jang, T. Y. (2005). Count data models for trip generation. *Journal of Transportation Engineering, 131,* 444-450.

Jung, B. C., Jhun, M., & Lee, J. W. (2005). Bootstrap tests for overdispersion in zero-inflated Poisson regression model. *Biometrics, 61,* 626-629.

King, G. (1989). Event count models for international relations: Generalizations and applications. *International Studies Quarterly, 33,* 123-147.

Lachenbruch, P. A. (2002). Analysis of data with excess zeroes. *Statistical Methods in Medical Research, 11,* 297-302.

Lachenbruch, P. A. (2001a). Comparisons of two-part models with competitors. *Statistics in Medicine, 20,* 1215-1234.

Lachenbruch, P. A. (2001b). Power and sample size requirements for two-part models *Statistics in Medicine, 20,* 1235-1238.

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

Lawal, B. (2003). *Categorical data analysis with SAS and SPSS applications*. Mahwah, NJ: Lawrence Erlbaum.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kunnert, P., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, P. (2005). Zero tolerance ecology: Improving ecological inference by modeling the source of zero observations. *Ecology Letters, 8,* 1235-1246.

Mazerolle, M. J. (2004). *Mouvements et reproduction des amphibiens et tourbiès perturbèes*. Unpublished doctoral dissertation,. Universitè Laval, Québec.

McCullagh, P., & Nelder, J. A. (1989). *Generlised Linear Models*, (2nd ed.). London: Chapman & Hall.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Min, Y. (2003). Modeling data with clumps. *Dissertation Abstracts International - B, 64(12).* (UMI3117356).

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

Mooney, C. Z. (1997). *Monte Carlo Simulation.* Thousand Oaks, CA: Sage.

Moulton, L. H., Curriero, F. C., & Barroso, P. F. (2002). Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research, 11,* 317-325.

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

Mwaliki, S. M. (January, 2005). The zicounts package. Retrieved May, 2006 from http://cran.r-project.org/doc/packages/zicounts.pdf

Pardoe, I., & Durham, C. A. (2003). Model choice applied to consumer preferences. *In Proceedings of the 2003 Joint Statistical Meetings,* Alexandria, VA, American Statistical Association.

Paxton, P., Curran, P. J., Bollen, K. A, Kirby, J, & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287-312.

R Development Core Team (2006). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis.* (2nd ed.). Belmont, CA: Wadsworth.

Ridout, M.S., Demétrio, C.G.B. and Hinde, J.P. (1998) Models for counts data with many zeroes. *Proceedings of the XIXth International Biometric Conference*, Cape Town, Invited Papers, pp. 179-192. Paper retrieved March 13, 2006 from http://www.kent.ac.uk/ims/personal/msr/zip1.html

SAS Institute Inc. (2000). *SAS/STAT User's Guide*, V.8, Cary, NC: SAS Institute Inc.

Scheaffer, R. L. (1995). *Introduction to probability and its applications*. (2nd ed.). Belmont, CA: Wadsworth.

Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis and Prevention, 29,* 829-837.

Singer, J. D., & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

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,* 1-9.

SPSS (2006). SPSS for Windows, Rel. 14.0. Chicago: SPSS, Inc.

Stokes, M. E., Davis, C. S., & Koch, G. G. (1991). *Categorical data analysis using the SAS system*. (2nd ed.). Cary, NC: SAS Institute, Inc.

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,* 1-9.

Tooze, J. A., Grunwald, G. K., & Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research, 11,* 341-355.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57,* 307-333.

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,* 275-289.

Weiller, H. (1964). A significance test for simultaneous quantal and quantitative responses. *Technometrics, 6,* 273-285.

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., & Lindenmayer, D. B. (1996). Modelling abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling, 88,* 297-308.

Zhou, X., & Tu, W. (1999). Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics, 55,* 645 – 651.

Zorn, C. J. W. (April 18-20, 1996). Evaluating zero-inflated and hurdle Poisson specifications. Midwest Political Science Association, 1-16.

BIOGRAPHICAL SKETCH

Jeffrey Monroe Miller was born in Greensboro, NC and raised in Jacksonville, FL. He earned a Bachelor of Arts degree from the University of Wisconsin Eau Claire in 2001 majoring in Psychology with two minors in Music. As an undergraduate, he became fascinated by the methods used to analyze data. In 2002, he enrolled in the Research and Evaluation Methodology program at the University of Florida in the Department of Educational Psychology (College of Education). In 2004, after authoring and defending a thesis pertaining to item-order effects in surveys, he received the Master of Arts in Education degree in this program. In 2007, he completed the requirements for the degree of Doctor of Philosophy.