Comparison of the Performance of Count Data Models under Different Zero-Inflation Scenarios Using Simulation Studies

Yuan Jiang (jyspring@ufl.edu), PhD. Student
Lisa House (lahouse@ufl.edu), Professor
Food and Resource Economics Department
University of Florida

Selected Paper prepared for presentation at the
2017 Agricultural & Applied Economics Association's Annual Meeting, Chicago, IL, July 30-August 1,2017

Comparison of the Performance of Count Data Models under Different Zero-Inflation Scenarios Using Simulation Studies

Abstract

When analyzing consumption behaviors on the individual level, data is often in the format of count data. A challenge with this data is that there are many zero observations (zero-inflated) because there may be many observations recorded as zero-consumption during a given period. In this paper, we aim to examine the performance of six count-data models under different zero-proportion, and skewness levels using simulation studies. Additionally, we further compare the capabilities of these models on predicting zero-observations, and structural zero-observations, in order to evaluate their capabilities in predicting market structure when applying to the food consumption analysis. Based on this study, it is recommended to the researchers to consider the hurdle models when there is zero-deflation, and the zero-inflated models when there is zero-inflation. If the underlying assumption assumes that there are different types of zero observations, it is recommended to use zero-inflated models.

Key word: Count Data Model, Simulation Studies, Food Consumption, Hurdle Model, Zero-inflated Model

Comparison of the Performance of Count Data Models under Different Zero-Inflation Scenarios Using Simulation Studies

Introduction

Agricultural economists have long been interested in modeling consumers' behaviors, in particular consumers' preferences and purchases. In models analyzing consumption behavior on an individual level, data is often in the format of count data. A challenge with this data is that there are many zero observations (zero-inflated) because there may be many observations recorded as zero-consumption during a given period.

In statistical modeling, when the dependent variable is count data, the most popular regression technique is the Poisson regression model. However, the Poisson model fails to provide an adequate fit when there exists the problem of zero-inflation. Thus, the Poisson model has been modified to solve this issue. The most popular modification is the Zero-Inflated/Modified Poisson model and Hurdle Poisson model. Further, there are negative binomial variations of these models considering the possible issue of overdispersion.

The zero-inflated Poisson (ZIP) model was proposed by Lambert in 1995. Following this, a number of related models have been proposed, including the Poisson-negative binomial and modified Poisson suggested to address inequality of the mean and variance (as equality is assumed for the Poisson distribution) (Famoye and Singh, 2006). The zero-inflated count data model assumes that the zero observations coming from two distinct sources: "sampling zeros" and "structured zeros." When applied to consumption analysis, zero-consumption could be recorded when the consumer is genuine non-participant (structure zero), or when the zero consumption is the corner solution of a standard consumer demand problem (sampling zero).

Different from the Zero-inflated count data model, the Hurdle models proposed by Mullahy (1986) assumes that all the zeros are sampling zeros. When applied to consumption analysis, it assumes that individuals need to pass two stages before being observed with a positive level of consumption: a participation decision and a consumption decision. Furthermore, the hurdle models assume participation dominant. Thus, all the zero observations are assumed generated in the first stage (whether to consume), and in the second stage, the consumption behavior is truncated at zero.

Thus, the choice between hurdle models and zero-inflated models is typically based on whether the researcher believes that all the zero observations are coming from the structural zero group or that at least some of the zeros are sampling random zeros.

There has been relatively little literature trying to compare and evaluate the performances of these count data models, and the results of the studies that do exist do not come to the same conclusion. For example, Green (1994) found that the negative binomial model was superior to the ZIP model, and the ZIP model was superior to the Poisson model; However Lambert (1992) argued that ZIP model had superior fit to Negative Binomial model. Needlon et.al.,(2010) found that the ZIP model fits better than the Poisson and Hurdle models, while Welsh et.al.(1996) found that the Hurdle and ZIP models to be equal. Based on Miller's (2007) research, the discrepant results of the model comparisons might be because the datasets they employ are quite different in the proportion of zeros, with some research using data with 20% zeros, and some datasets with as much as 90% zeros.

In addition to differently structured data with respect to zeros, there is also difference in datasets with regards to overdispersion. Few studies have used simulated data to examine one or both of these issues. Lambert (1992) proposed the zero-inflated Poisson model and evaluates its

performance using simulation studies. Miller (2007) compares the Poisson, hurdle, and zero-inflated models under varying zero-inflation levels; Desjardins (2013) evaluates the performance of zero-inflated negative binomial and negative binomial hurdle models under simulation. However, most of these previous studies compare the models mainly focus on its model fits and parameter recovery. However, how well the models predict different categories of consumers can also be of importance, especially for consumption studies. When choosing between the hurdle models which assume no structural zeros, and zero-inflated models, which allow both structural and sampling zeros, what are their capabilities of predicting market segmentation is of interest.

To the best of our knowledge, there has been no simulation studies conducted that compare zero-inflated and hurdle models with both Poisson distribution and Negative binomial distributions. There are no prior studies investigating how zero proportions and levels of overdispersion might affect estimations and model fit in the hurdle models and zero-inflated models. Furthermore, special attention can be given to the comparison of model capabilities to predict the correct latent classes. In the case of consumption analysis, we will compare the models' capabilities of predicting the market structure and segmentations. Considering the assumption that the zero-inflated models have two different types of zeros, yet the hurdle models only assume the existence of one type of zero. It would be very important to test whether the zero-inflated models would efficiently predict the different types of zeros, especially given the different levels of zero portions.

In this study, we will examine two research questions. First, under different levels of zero proportions, and skewness of positive outcomes, how will these count data models (Poisson model, Zero-Inflated Poisson model, Hurdle Poisson model, and their negative binomial variations) perform in terms of model fit. Secondly, we will pay special attention to the

comparison between Zero-Inflated Models and Hurdle models, and evaluate the proportion of correctly identified structural zeros for the Zero-Inflated Models, and test the consequences, if any, of misspecifying the latent classes for the zeros given different levels of structural zeros.

Literature Review

Count Data and Generalized Linear Model

Count data occurs very frequently in many different fields of research, especially in the field of social science. Count data can be used to represent the number of times that an event occurs under a certain condition or during a certain time; for example, the number of times that consumers purchase a certain good during a certain period would be an event count. As such, the response values take the form of discrete non-positive integers. Hence, count data is the "realization of a nonnegative integer-valued random variable" (Cameron and Travedi, 1998)

When analyzing count data, there is an assumption that the number of events is independently identically distributed with a discrete probability distribution. The most common probability distributions used to describe count data are the Poisson and Negative Binomial distributions. The Poisson distribution was derived by Poisson (1837) as a limiting case of the binomial distribution, with the characteristics of mean-variance equality. The negative binomial distribution was derived by Greenwood and Yule (1920) and was used as an alternative to the Poisson distribution when violating the mean-variance equality.

As for the regression models, the classic linear regression model is not suitable for count data analysis, since the assumption of normality is violated. Thus, generalized linear models,

which allow the analysis of data when the assumptions of linearity and normality are no longer met, are employed.

The Generalized Linear Model (GLM) was first described by Nelder and Wedderburn (1972) and has been further developed and explained by McCullagh and Nelder (1989). Instead of modeling the mean as a linear function of the covariance in the classic linear regression, it allows other possibilities. All GLM are specified with three components: a random component which specifies the distribution of the output variable; a systematic component which specifies the covariants in a linear form and a link function which connects the random component to the systematic components. If the distribution of the output variable is normal, then it is the classic OLS regression. Besides the normal distribution, other distributions, like binomial distributions, Poisson distribution, negative binomial distributions, etc. can be used.

To be more specific on the three components of the GLM, it is necessary to clarify the equations. The systematic component is of the linear form of the covariants as follows:

$$\eta = x'\beta$$

Where x_i is the vector of covariance for observation i, and β is the correponding unknown parameters.

A link function connects the mean value of the output variable Y to the linear predictor η through a function $g(.)$.Thus, the GLM model is expressed as

$$g(\mu) = \eta = x'\beta$$

Poisson Regression Model and Applications

The Poisson regression model is the most popular method for analyzing count data. It is a specific form of the GLM which specifies the output variable Y being followed by a Poisson distribution, with the link function $g(\mu) = \log(\mu)$.

Thus, the probabilities of observing y_i can be written as：

$$f(y_i, \mu_i) = \frac{\mu_i{}^{y_i} e^{-\mu_i}}{y_i!} \quad y_i = 0,1,2,3 \ldots\ldots$$

Where μ_i is parameter of the Poisson regression, which is also the mean and variance of y_i for the ith observation. Given the link function and the linear predictor η_i $g(\mu_i) = \log(\mu_i) = x_i{}'\beta$, thus we have $\mu_i = \exp(x_i'\beta)$.

The Poisson regression has been widely used when analyzing count data. In the field of consumption behavior, because the consumption frequency or purchase intensity is often described as count data, Poisson regression models have been used to analyze consumers' behavior. For example Morland et al. (2002) used the Poisson regression when analyzing consumers' access to healthy food choices concerning to the distribution of food stores and food service places. Binkley (2006) employed the Poisson regression to explore the effect of demographic, economics and nutrition factors on the consumption frequency of food away from home. Cannuscio et al. (2013) using the Poisson regression to analyze the correlation between food environment and residents' shopping behaviors.

Problems with the Poisson Regression

Although Poisson regression models are popular when analyzing count data, the model might not be the best fit due to the characteristic of the Poisson regression of the mean-variance equation, which is specified by $\mu_i = E(Y_i) = Var(Y_i)$. The assumption is very restrictive and is

easily violated. When the observed variability is greater (or less) than the observed mean, then the Poisson distribution would no longer be the true realization of the data, and the data is considered to have the issue of overdispersion (underdispersion). Taking the case of consumption behavior as an example, for some daily goods like tobacco, there might be many people choose to never consume tobacco because they are non-smokers, yet there might be also many people choose to consume extremely large units per week (heavy-smokers). In this case, the data might not meet the assumption of mean-variance equality, and Poisson model would not be appropriate.

A special case of overdispersion happens when there are excessive zeros in the data. When there are abundant zeros, the mean of the data will be closer to resulting in the violation of mean-variance equality assumption. Thus, ignoring the issue of excessive zeros will cause biased parameter estimates and poor model fit. Using tobacco consumption as an example again, whenever the question of consumption frequency is asked, there would be many people who answer zero, since they are non-smokers. The same thing happens for the consumption of food, where consumers might choose not to consume in a given time period, or not to consume for reasons such as allergies or personal beliefs.

When considering the source of the excessive zeros, some research argues that the zeros might arise from different generating processes, which is a result of unexplained population heterogeneity (Hu et al.,2011; Rothman 2013). Generally, it was considered that the zeros could be differentiated into two types: structural zeros, which are generated from a latent class that zero is the only possible value, and sampling zeros, which arise from a latent class where zero happens within a random sample of potential count responses. In the case of consumption behaviors, in response to the question “How often did you consume seafood last month” there

will be individuals who never consume seafood before (structural zeros) and individuals who are potential consumers but might not choose to consume in the last month(structural zeros). The prior structural zero observations are the consumers who have a non-positive desire (which can be categorized as non-participants), and the posterior sampling zero observations are those consumers who have positive desire but no positive acquisition in the given period (which can be categorized as potential consumers).

To deal with the issues of overdispersion and excessive zeros, different models have been used. Generally, when overdispersion is the only issue, the negative binomial model will be a better fit than Poisson regression. If only zero-inflation exists, either the Zero-inflated Poisson model or hurdle Poisson model are used. If both exist, then zero-inflated negative binomial and hurdle negative binomial models could be used.

Negative Binomial Regression Models and Applications

When the data has the issue of overdispersion, the negative binomial model is usually considered as an alternative to the Poisson regression model, since it provides an extra parameter to accommodate the additional variability. The negative binomial distribution (NB) is a gamma mixture of the Poisson distribution. In other words, a random non-negative integer is considered distributed as the Poisson distribution with a mean of λ, where λ is a random variable with a gamma distribution. Thus, the NB works to allow more flexibility in accommodating variability.

For example, for the gamma distribution with shape parameter γ, and scale parameter $\theta=\frac{\rho}{1-\rho}$, the mass function of the negative binomial distribution given the gamma-Poisson mixture is written as:

$$f(y,\gamma,\rho) = \int_0^{\infty} f_{poi(\lambda)} * f_{gamma(\gamma,\rho)}(\lambda)\, d\lambda$$

$$= \int_0^{\infty} \frac{\lambda^y}{y!} e^{-\lambda} * \lambda^{r-1} \frac{e^{-\lambda\frac{\rho}{1-\rho}}}{\left(\frac{\rho}{1-\rho}\right)^{\gamma}\Gamma(\gamma)} d\lambda$$

$$= \frac{\Gamma(\gamma+y)}{y!\Gamma(\gamma)} \rho^y (1-\rho)^{\gamma}$$

The standard formulation for the negative binomial mass function of a variable Y is given in the following form:

$$f(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)} \left(\frac{k}{k+\mu}\right)^k \left(1 - \frac{k}{k+\mu}\right)^y$$

where $E(Y) = \mu$, and $Var(Y) = \mu + \frac{\mu^2}{k}$, and $\frac{1}{k}$ is defined as the dispersion parameter. k is the gamma scale parameter. As k increases to infinity, Var(Y) decreases to u, which is equal to E(Y), and the distribution of negative binomial approaches the Poisson distribution.

For the negative binomial regression model, which is also a specific form of GLM, the link function is also the log transformation as the Poisson regression model $g(\mu_i) = \log(\mu_i) = x_i'\beta$. Further more, as mentioned above, the negative binomial distribtuion converges to the poisson distribution if k increases to infinity, thus, the Poisson regression model is nested within the negative binomial regression model. As a result, the Likelihood Ratio Test or Wald test can be used to test whether the dispersion parameter is significant.

Since negative-binomial regression models are more flexible than Poisson regression models accommodating data with more variability, it has also been widely used in the analysis of consumption behavior. For example, Lesser et al (2013) employ the negative binomial model to test the association between outdoor food advertising and obesity. When analyzing the data, the authors reject the Poisson model because of the existence of overdispersion. Han and Powell

(2013) also employed the negative binomial model to analyze consumption patterns of sugar-sweetened beverages in the United States.

However, although the negative-binomial regression models can accommodate issues of data with overdispersion, these models also have some limitations, especially when dealing with zero-inflation. Previous research indicates that negative binomial regression models are not a good fit for data with zero-inflation (Desjardins, 2012; Hu et al., 2011; Lambert, 1992). Additionally, considering the potential different latent classes which generate two types of zeros, using the negative binomial models could not capture the different characteristics of the different groups.

In the case of consumption, negative-binomial models would be very restrictive by assuming that it is the same set of factors that influence both consumers' decisions on participation and consumption. Furthermore, both Poisson regression models and negative binomial regression models assume that the characteristics of consumers and non-consumers have no significant difference, thus fail to identify different consumer types. To investigate the different types of zeros (and consumers) , a mixture model or a two-part model may improve fit.

Zero-inflated Models

Zero-inflated models refer to models that define a mixture of two different distributions at zero, and are able to accommodate the issue of excessive zeros in count data. Zero-inflated models assume that there are different latent classes in the population. Thus, the zero observations could be generated through two different sources: "sampling" and "structured" zeros. When applying the zero-inflated model to the consumption case, the observed zero consumption will be recorded when the consumer is genuine non-participant (structured zero), or

when the consumers are potential consumers, and choose zero consumption as the corner solution of a standard consumer demand problem (sampling zero). Thus, using the zero-inflated models will allow us to predict the existence of three different groups: genuine non-participant, potential consumers, and active consumers with positive consumption.

Zero-inflated models have been developed for different models, including Poisson regression models (Lambert, 1998), negative binomial regression models (Ridout, Hinde, and Demetrio, 2001), and other models (geometric models (Mullahy, 1986)).

The zero-inflated Poisson (ZIP) model was proposed by Lambert in 1995. It assumes a mixture of two distributions at the point of zero: a Poisson distribution and a binomial distribution. According to this assumption, it is assumed that with probability p, the only possible observation is 0 (structural zeros), and with probability (1-p), a Poisson random variable is observed. The probability mass function of a ZIP model is as follows:

$$\Pr(Y=y)=\begin{cases} p+(1-p)\exp(-\lambda)\, I_{(y=0)} \\ (1-p)\frac{\lambda^{y}e^{-\lambda}}{y!}\, I_{(y>0)} \end{cases}$$

Thus, from the above equation, zero observations can be observed from two parts: structural point mass component , p, and from the sampling Poisson component, $(1-p)\exp(-\lambda)$. In the ZIP model, $E(Y)= \mu = (1-p)\lambda$, and $Var(Y)= \mu + \frac{p}{1-p}\mu^2$.

The ZIP model is also a special case of GLM, with a logit link function for p, and log link function for λ as follows:

$$\text{Logit}(p)=\text{Log}\left(\frac{p}{1-p}\right)=x'\beta$$

$$\text{Log}(\lambda) = z'\alpha$$

Where x are the covariates for the first stage, with β as the corresponding estimates; z are the covariates for the second stage, with α as the corresponding estimate. Furthermore, this is no requirement that x=z.

The zero-inflated Poisson model has been widely used when dealing with excessive zeros, and there are many examples in the analysis of consumption. For example, Almasi et al. (2016) employed the ZIP model to analyze the effects of nutritional habits on dental care among schoolchildren. Matheson et al. (2012) explored the influence of gender and neighborhood deprivation on alcohol consumption using the ZIP model.

Additionally, Lambert (1992) also extends the ZIP model to the ZIP(τ) model which allows p and λ to be correlated with a shape parameter τ. Huang and Chin (2010) employed the ZIP(τ) to model road traffic crashes. Calsy et al. (2009) explored the correlation with motivational and skill training and HIV transmitted infection sexual risk using the ZIP(τ).

Just as that the Poisson regression model was extended to the negative binomial regression model, the zero-inflated Poisson regression model can also be extended to the zero-inflated negative binomial regression model. Even with zero-inflation, it is also possible that overdispersion happens because of the greater variability of the non-zero outcome. In this case, instead of the ZIP model, the ZINB model is a better fit for the data.

Similar to the ZIP distribution, the ZINB distribution assumes that there is a mixture distribution at the point of zero: a negative binomial distribution and a binomial distribution. Thus, the ZINB can be expressed as follows:

$$\Pr(Y=y)=\begin{cases} p+(1-p)\left(\frac{k}{k+\mu}\right)^{k} \; I_{(y=0)} \\ (1-p)\frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}\left(\frac{k}{k+\mu}\right)^{k}\left(1-\frac{k}{k+\mu}\right)^{y} I_{(y>0)} \end{cases}$$

Where u is the mean of the NB distribution, and $\frac{1}{k}$ is the dispersion parameter. Thus the mean and variance of the ZINB distribution is E(Y)= $(1-p)\mu$， Var(Y)= $(1-p)*\mu*(1+\frac{\mu}{k}+p\mu)$. Just as the NB distribution converges to the Poisson distribution as k increases to infinity, the ZINB distribution also converges to the ZIP distribution as k increases. Examples of the use of the ZINB model in consumption include Hendrix and Haggard (2015), who employed the ZINB model to analyze global food prices and regime type in the developing world.

The Hurdle Model

The hurdle model was first developed by Cragg (1971) as an example of truncated models, relaxing the Tobit model by allowing separate stochastic processes for the observed zeros and positive outcomes (Yen and Huang, 1996). Different from the zero-inflated models, the hurdle models are no longer a mixture of different models, but a two-part model. The first part predicts whether the outcome is zero or not, and the second part generates the non-zero counts. Thus, it assumes that all the zeros are from the first stage.

When modeling consumption behavior using the hurdle count data model, there is an assumption that individuals need to pass two stages before being observed with a positive level of consumption: a participation decision and a consumption decision. In the first stage, the consumers make decisions on whether to participate or not. In the second stage, a decision on how much/many to purchase was determined. Specifically, the hurdle model assumes that the participation stage dominates the consumption stage. Thus, if the consumers choose to participate in the first step, it does not allow zero-consumption in the consumption stage.

In prior research, a hurdle model uses a binomial logistic regression model to indicate whether a count is zero or positive (Green, 1994). If a positive outcome is realized, then it a

truncated at zero count data model (Poisson/NB) is used for the positive counts. However, the first part does not have to be the binomial logistic regression model, "there will likely exist numerous plausible specifications of both the binary probability model and the conditional distribution of the positives" (Mullahy, 1986). For example, in Mullahy's research (1986), he used the Poisson distribution governing the probability of observing a zero count. Thus, a generic hurdle model is as follows:

$$\Pr(Y=y)=\begin{cases} g_1(0)\, I_{(y=0)} \\ \left(1-g_1(0)\right) * \frac{g_2(y)}{1-g_2(0)}\; I_{(y=1,2,3\ldots.)} \end{cases}$$

Where Y is the outcome variable, g_1 is the specification of the binary probability model that governs the first hurdle, indicating whether the outcome is zero; g_2 is the specification of the trucated-at-zero probability generating the positives.

There are also some popular specifications for g_1 and g_2, for example Green (1994) specified the g_1 as a binomial distribution and g_2 as a truncated-at-zero Poisson distribution, which provides the following form:

$$\Pr(Y=y)=\begin{cases} p\, I_{(y=0)} \\ (1-p) * \frac{\lambda^y e^{-\lambda}}{(1-e^{-\lambda})y!}\; I_{(y=1,2,3\ldots.)} \end{cases}$$

Where p is the probability of a count being observed as zero, and λ is the parameter for the truncated Poisson distribution. To be more specific, the link function for p is logit transformation where Logit (p)=Log($\frac{p}{1-p}$)=$x'\beta$, and the link function for λ is log, with Log(λ) = $z'\alpha$. x are the covariates for the first stage, with β as the corresponding estimates; z are the covariates for the second stage, with α as the corresponding estimate. Furthermore, this is no requirement that x=z.

Mullay (1986) specified both g_1 and g_2 as Poisson distributions which provides the following specifications:

$$\Pr(Y=y)=\begin{cases} e^{-\lambda_1} I_{(y=0)} \\ \left(1-e^{-\lambda_1}\right) * \frac{{\lambda_2}^y e^{-\lambda_2}}{(1-e^{-\lambda_2})y!} \; I_{(y=1,2,3....)} \end{cases}$$

Where λ_1 is the parameter for the Poisson distribution governing the first hurdle; λ_2 is the parameter for the Poisson distribution generating the positives. Both λ_1 and λ_2 could be parameterized with log link function as $\text{Log}(\lambda_1) = x'\beta$, and $\text{Log}(\lambda_2) = z'\alpha$.

Shonkwiler and Shaw (1996) extended Mullahy's specification by allowing zero observations in both the first and second stage. Thus, in Shonkwiler and Shaw's model (Double hurdle count-data model[1]), there are two mechanisms generating zero observations: zero observations could either happen in the first stage by choosing not consume or in the second stage by choosing consume zero frequency. However, the essence of the double hurdle count data model is very similar to the ZIP model, but with the first part indicating the structural zeros using Poisson distribution specification instead of a binomial. The specification for the double-hurdle count data model is as follows:

$$\Pr(Y=y)=\begin{cases} e^{-\lambda_1} + \left(1-e^{-\lambda_1}\right) * e^{-\lambda_2} I_{(y=0)} \\ \left(1-e^{-\lambda_1}\right) * (1-e^{-\lambda_2}) \frac{{\lambda_2}^y e^{-\lambda_2}}{(1-e^{-\lambda_2})y!} \; I_{(y=1,2,3....)} \end{cases}$$

$$= \begin{cases} e^{-\lambda_1} + \left(1-e^{-\lambda_1}\right) * e^{-\lambda_2} I_{(y=0)} \\ \left(1-e^{-\lambda_1}\right) * \frac{{\lambda_2}^y e^{-\lambda_2}}{y!} \; I_{(y=1,2,3....)} \end{cases}$$

Where λ_1 is the parameter for the Poisson distribution governing the first part, indicating whether the zeros are structural zeros or not; λ_2 is the parameter for the Poisson distribution for the second part. Both λ_1 and λ_2 could be parameterized with log link function as $\text{Log}(\lambda_1) = x'\beta$,

[1] The term borrowed from Shonkwiler and Shaw (1996)

and $\text{Log}(\lambda_2) = \text{z}'\alpha$. If we let $\text{p} = e^{-\lambda_1}$, then this model specification is the same as the ZIP model.

The Poisson regression model can be extended to NB regression model, and the ZIP model can be extended to ZINB model, the hurdle Poisson regression model can be extended to the hurdle NB model. Examples of research using hurdle models include Crowley, Eakins and Jordan (2012), who employed the double-hurdle model to analyze the lottery participation and expenditure; Jaunky and Ramchurn (2014), who analyzed consumer behavior in the scratch card market using a double-hurdle model; Jiang and Lisa (2012) used double-hurdle approach to model mushroom consumption, and Bezu and Kassie (2014), who used the double hurdle model to estimate maize planting decisions.

Comparison of the Models

In this section, there are 6 count data models listed, including the Poisson, NB, ZIP, ZINB, Hurdle Poisson, and Hurdle NB models. When models are nested within one another, a Wald/LR test can be used to test the significance of these extra parameters. For example, the Poisson regression model is nested within NB, the Poisson Hurdle model, and the NB hurdle models; The ZIP is nested within ZINB; the PH is nested within the NBH. Besides these, the other pairs of the models are not inherently nested within each other. If models are not nested, they can be compared using Vuong's test, AIC and BIC.

Prior research has compared some models, such as the NB and Poisson regression models, where research, as mentioned above, showed the NB better handles the problem of overdispersion (Atkins and Gallops, 2007; Warton, 2005). According to Warton's research

(2005), they also found that when the overdispersion was not present, Poisson regression models performs better than the NB model.

As for the comparisons between NB and Zero-inflated models, Lambert (1992) compared the ZIP model to the Poisson and NB model when the ZIP model was first proposed. The conclusion was that the ZIP outperformed the other two models, and NB performs better than Poisson model. Green (1994) compared the NB, ZIP, and ZINB models. Based on the Vuong test statistics, he found that the ZINB model performs the best, followed by the NB, ZIP and Poisson models. A possible reason for this result may be because the ZINB model could accommodate two sources of overdispersion, and in the data used in this study, the overdispersion was caused mostly by unobserved response heterogeneity. This would lead the NB model to perform better than the ZIP model. Desouhant et al. (1998) compare the NB and the ZIP model and found that the two models perform roughly similar, and they conclude that researchers need to accommodate both overdispersion and zero-inflation in the analysis of count data. Slymen et al. (2006) compared the ZIP, ZINB, NB, and Poisson models and found the NB model fit better than the Poisson models. However, the ZINB and ZIP models performed nearly same both in terms of model fit and parameter estimates, which indicates that the main issue of the data in this study was zero-inflation, and the overdispersion is not severe in this case. Wenger and Freeman (2008) compared the ZIP, ZINB, NB, Poisson and concluded that the zero-inflated models perform better than the non-inflated models; the NB formulation models fit better than other models without the NB formulation.

The comparisons between the Zero-inflated models and the Hurdle models has driven more attention, focusing on two main differences between models. At first, the hurdle models assume there only exist one type of zero observations, yet the zero-inflated models assume that zero

observations are coming from two different sources. Second, zero-inflated models are typically used to analyze data with zero-inflation and have a poor fit for data with under-dispersion of zero counts, yet the hurdle models have better fit dealing with zero-deflation. Min and Agresti (2005) compared the hurdle models, and zero-inflated models, and found the PH model had a better fit than the ZIP. Desjardins (2013) compared the ZINB and HNB models using simulations and found that the HNB performs better than the ZINB in terms of both model fit and parameter recovery.

Gaps and Shortcomings

Although there is much research comparing models, there has been very few studies comparing and evaluating the model performance using simulation studies. Lambert (1992) used simulation to compare the ZIP model with the Poisson and NB model the ZIP model was first proposed. Miller (2007) compared the performance of the count-data models under different levels of zero-proportions and skewness using simulation. This research is very interesting, yet it still has several limitations. At first, when comparing the model with NB formulations to those without NB formulations, different levels of overdispersion are not accounted for. Next, when comparing the models, the author focuses mainly on model fit, without consideration of prediction. More recently, Desjardins (2013) compared the ZINB and HNB model under simulation allowing different levels of overdispersion, and sample size. However, in this research, it is only two models being compared with the pre-assumed portion of zeros.

Thus, an area that needs further investigation is how the models perform given different levels of zero-proportion and overdispersion. From the previous empirical research, we can infer that if the effect of overdispersion is much larger than the effect of zero-inflation, the NB model

should perform better than the ZIP model. The question remains, how does the effect of zero-inflation change given different levels of overdispersion? How would the model performance change based on different levels of zero proportion given different levels of dispersion?

Furthermore, when analyzing consumption behavior, we are extremely interested in analyzing different consumer types and exploring the market structure and segmentation. In this sense, instead of the model fit, we also care about the capability of prediction for the models.Yet, there has been no prior research focusing on prediction capabilities. Especially when comparing the Zero-inflated models to the hurdle models, the latter allows two different types of zeros, and the former only allows a single type of zero observations. So it would be very interesting to see if the zero-inflated model would efficiently predict the correct portion of non-consumers (structural zeros), which is the most important utility of the zero-inflated models. With one more step, if we allow different levels of structural zeros in the Zero-inflated models, how would the performance of the two models change?

Method

Research Questions

Based on the literature review, we will examine two research questions in this study. First, under different levels of zero proportions, how will these count data models (Poisson model, Zero-Inflated Poisson model, Hurdle Poisson model, and their negative binomial variations) perform regarding model fit. Secondly, we will pay special attention to the comparison between Zero-Inflated Models and Hurdle models, and evaluate the proportion of correctly identified structural zeros for the Zero-Inflated Models, and test the consequences, if any, of misspecifying the latent classes for the zeros given different levels of structural zeros.

To answer the first research question, a simulation study was conducted. We employ a Monte Carlo design to sample 1,000 cases based on different levels of zero proportions and skewness of non-zero counts. In this study, the zero-proportion and skewness will be varied. For each dataset, each model will be analyzed with 2,000 simulations, and then deviance statistics and AIC values will be used as the measurement to compare the model fit.

To answer the second research question, a second simulation study was conducted when the true model was known. We will generate datasets from four different distributions: Zero-Inflated Poisson distribution ($\rho,\ \mu$) (where ρ is the proportion of structural zeros, and μ is the mean of Poisson); Hurdle Poisson (π, γ), Zero-Inflated Negative Binomial distribution (($\rho,\ \mu,\ k$) where ρ is the proportion of structural zeros, μ is the mean of Poisson, k is the dispersion rate), and Hurdle Negative Binomial distribution (π, γ, k), and fit each datasets with each of the six different count-data models to compare their performances under different true model specifications. Specifically, we can set the zero/structural zero percentages at different levels, and we can also control the different levels of k (over-dispersion), and compare their capabilities of capture the zero observations, and structural zero observations.

Monte Carlo Models

As discussed in the previous section, the generalized linear model was constructed by a systematic component, a random component and a link function.The case model for the Poisson regression assumes that $y_1, y_2 \ldots y_n$ are independently dependent following the distribution:

Y_i~Poisson (θ_i)

Where the link function is

$\log(\theta_i) = \beta_0 + \beta_1 * (x_{1i}) + \beta_2 * (x_{2i})$

Similarly, the negative binomial formulation of the Poisson model is the same as the Poisson regression but with an extra parameter of dispersion.

The case model for zero-inflated poission model assumes that $y_1, y_2 \ldots y_n$ are independently dependent following the distribution:

Y_i~ZIP (p_i, θ_i)

Where the link function is

Logit (p_i)=Log$(\frac{p_i}{1-p_i})$=α_0+$\alpha_1 * (z_{1i})$+ $\alpha_2 * (z_{2i})$

log(θ_i)= β_0+$\beta_1 * (x_{1i})$+ $\beta_2 * (x_{2i})$

The zero-inflated negative binomial model is similar to the ZIP model, with an extra overdispersion parameter as Y_i~ZINB (p_i, θ_i, k^{-1}), the link function of ZINB model is the same as the ZIP model.

The last set of models are the Hurdle models. The Hurdle Poisson regression model assumes that $y_1, y_2 \ldots y_n$ are independently dependent followed the distribution:

Y_i~HP (π_i, θ_i)

Where the link function is

Logit (π_i)=Log$(\frac{\pi_i}{1-\pi_i})$=α_0+$\alpha_1 * (z_{1i})$+ $\alpha_2 * (z_{2i})$

log(θ_i)= β_0+$\beta_1 * (x_{1i})$+ $\beta_2 * (x_{2i})$

Similarly, the Hurdle Negative Binomial model has the same link function as the HP model, but has one addition parameter for overdispersion.

First Simulation Study – Model Comparison for the six models in terms of model fit

The first simulation study was designed to examine the performance of the six count data models under different levels of zero-proportion and skewness levels. In this experiment, we assume that there are five levels of zero-proportion: 0.1, 0.3, 0.5, 0.7 and 0.9 (i.e. 0.1 = 10% zeros in the dataset). As for the skewness, we assume three levels: positive-skewness (skewness=1.3), normal (skewness=0), and negative-skewness (skewness=-1.3).

The sample size is an important concern when analyzing different models. When the sample size is too small, results may not be consistent since it is not valid to assume asymptotically normal, however, when the sample size is too large, the computation time is significantly increased. Thus, in this study, we choose the sample size n=1,000, which was chosen based on the prior research.

The simulation size is also important considering the validation of simulation results. If there are too few replications, the results would be not consistent, and as the size of simulation is increasing, the consistency of the results would also increase. In this study, we choose the simulation size equal to 2,000, similar to Lambert (1992) which has been proven to produce asymptotic results.

Data generating process for the first simulation study is described below. Given different levels of zero-proportion, the total number of counts to be sampled was one minus the prespecified zero-proportion. For example, if the zero-proportion is 0.5, it means that we need to randomly draw 500 count number (from 1,2,3,4,5). Then, given the three different levels of skewness, count number was randomly drawn using the *"sample"* procedure in R.

As for the process of analysis, the "*glm*" procedure in R was used for the Poisson and Negative Binomial Regression analysis, and "*pscl*" library in R (authored by Simon Jackman)

was used for the zero-inflated and hurdle models. For each model regressed on each dataset, results including log-likelihood, AIC, and coefficient estimates are saved for the further analysis.

Second Simulation Study – Model Comparison between hurdle and zero-inflated models

The second simulation study was designed to compare zero-inflated models with hurdle models given the set of simulation conditions. In particular, in this experiment we will test the performance of the the models given different levels of structural zero/zero proportions (0.1 ,0.3 ,0.5 ,0.7).

In this experiment,we will generate datasets from four different distributions: zero-inflated poisson distribution, negative binomial distribution, hurdle poisson distribution, and hurdle negative binomial distribution. For each distribution, the zero/structural zero proportion which is generated from a binomial process is controlled by different levels of p values, and the count process (Poisson/negative binomial) will be set given known coefficients. As for negative-binomial formulations, we also controlled different levels of overdispersion in order to compare the model performances under different situations. Then, for each type of dataset, we run six different count data models, and compare model fit, with a focus on their capabilities to capture zero and structural zeros. We will also evaluate their capability of coefficient recovery (for the counting process), and relative/absolute bias.

What is more, the previous research indicates that the model fit and their capabilities of capturing zero-observations would change given different sample size, thus, in this experiment, we also control the sample size at different levels. Based on the previous research, Lambert (1992) considered the sample size ranged from 25, 50, and 100, but sigularities and non-convergence happened in her experiment. Thus, similar to the experiment conducted by

Desjardins(2013), in order to avoid these issues, I increased the sample size to 100,200, and 500, and will also compare the model performances given different levels of sample size.

Model Evaluation of the Simulation Studies

In order to assess the performance of the six different models, we will employ various measures related to model fit and parameter behaviors. Specifically, for model fit, we will employ the AIC statistics, and bias for the E(Y|X) and Pr(y=0), and also the capabilities of capture zero /structural zeros. As for the parameter recovery, since we only control the parameters for the counting process, we will use both the estimated parameters and their confidence interval coverage to evaluate how close of the estimated parameters to the true parameters. Since, in the second simulation study, we generate four different distributions, and for each distribution, we will run all the six count-data model. Thus, we would give special attention to the performances of models as a consequence of using a wrong model given true distributions.

Preliminary Results[2]

First-Simulation Findings

First, we focus on the results from the first simulation experiment. When the distribution is normal distributed, we could find that as zero-proportion goes up, all of the six models turns to perform better. Specifically, when the zero-proportion is only 10%, all the six models turns similar AIC statistics, while as zero-proportion increasing, we could observe that the performances of modified Poisson regression models get significantly better than Poisson

[2] This analysis is still ongoing; the results are preliminary. The full version will update later.

models. In particular, when the zero-proportion is as high as 90%, both the Zero-inflated and hurdle models return better model fit than NB and Poisson regression.

Furthermore, if we focus on the comparison between zero-inflated and hurdle models, we could see that in this case, Hurdle-Poisson model performs better than zero-inflated Poisson model when the zero-proportion is small. Thus we could see that hurdle model could handle the situation of "zero-deflation," while zero-inflated models could not. Another finding is that we find the negative binomial formulations have larger AIC statistics compared to the Poisson formulation, the reason might because that the negative binomial formulations have one extra parameter(size) than poison formulations, which works as the penalty causing the AIC statistics get larger .

Table: Mean AIC comparing six models with normally distributed datasets

Zero-Proportion	Poisson	NB	ZIP	ZINB	HP	HNB
10%	3585.21	3587.22	3570.34	3572.34	**3570.29**	3572.29
25%	3808.69	3793.78	3561.03	3563.03	**3561.02**	3563.02
50%	3754.84	3364.35	3011.97	3013.97	**3011.96**	3013.96
75%	2914.66	2169.25	**1942.16**	1944.16	1942.17	1944.17
90%	1717.94	1090.41	**982.60**	984.60	**982.60**	984.60

When the distribution is positively distributed, results are similar as before. We could find that as zero-proportion goes up, all of the six models turns to perform better. Specifically, when

the zero-proportion is only 10%, all the six models turns similar AIC, while as zero-proportion increasing, we could observe that the performance of modified poisson regression models are significantly better than Poisson models. In particular, when the zero-proportion is as high as 90%, both the Zero-inflated and hurdle models return better model fit than NB and Poisson regression.

Hurdle -models again handled the zero-deflation situations better than zero-inflated models, and when the zero-proportion gets larger, the performances of zero-inflated models get better. Although the data is positively skewed, we still did not find the advantages of using negative-binomial formulations.

Table: Mean AIC comparing six models with positive distributed datasets

Zero-Proportion	Poisson	NB	ZIP	ZINB	HP	HNB
10%	3586.17	3588.18	3571.48	3573.43	**3571.43**	3573.48
25%	3808.62	3794.55	3559.49	3561.49	**3559.48**	3561.49
50%	3752.35	3362.43	**3012.17**	3014.17	**3012.17**	3014.17
75%	2918.74	2170.57	1942.39	1944.39	1942.39	1944.39
90%	1717.05	1081.12	**982.65**	984.65	982.66	984.66

When the data is negatively skewed, we could see that when the zero observations are inflated in the datasets, Poisson regression model is the worst fit, and zero-inflated or hurdle models are much better than it. In this case we found that the zero-inflated poisson regression model returns the best model fit in all of the five situations of zero-proportion, one reson might because the data is left skewed, causing the mean value moving to the left (pulled to the zero).

Table: Mean AIC copmaring six models with positive distributed datasets

Zero-Proportion	Poisson	NB	ZIP	ZINB	HP	HNB
10%	3588.28	3590.29	**3573.02**	3575.02	3573.08	3575.08
25%	3811.68	3796.49	**3562.86**	3564.86	**3562.86**	3564.86
50%	3760.90	3366.55	**3015.23**	3017.24	3015.24	3017.24
75%	2921.83	2171.69	**1942.89**	1944.89	**1942.89**	1944.89
90%	1714.56	1080.78	**982.27**	984.27	982.28	984.28

I also pulled out the mean AIC statistics for six models at the three different levels of skewness. Overall, in the following five tables, we could observe that the zero-inflated and hurdle models handled the skewness better than Poisson and Negative Binomial regression, and what's more, we also found that Poisson regression performs the worst in each of the situation. Besides, we also found that the larger the zero-inflation rate, the larger difference was captured the negative binomial regression and Poisson regression. When the zero proportion is only 10%, we could not observe significant differences among the six different models.

What is more, when we observe more in detail on the hurdle models, and zero-inflated models, we could see that at 10% and 25% level zero observations, the Hurdle model handled better when the datasets are normal/positively distributed, however, when the data is negatively

skewed, zero-inflated model turns better model fit. While as the zero-proportion getting bigger, we found that zero-inflated models' performances on the normal/positive distributed datasets also gets better.

Table: Mean AIC comparing six models with 10% zero proportion

Skewness	Poisson	NB	ZIP	ZINB	HP	HNB
normal	3585.21	3587.22	3570.34	3572.34	**3570.29**	3572.29
positive	3586.17	3588.18	**3571.43**	3573.43	**3571.43**	3573.48
negative	3588.28	3590.29	**3573.02**	3575.02	3573.08	3575.08

Table: Mean AIC comparing six models with 25% zero proportion

Skewness	Poisson	NB	ZIP	ZINB	HP	HNB
normal	3808.69	3793.78	3561.03	3563.03	**3561.02**	3563.02
positive	3808.62	3794.55	3559.49	3561.49	**3559.48**	3561.49
negative	3811.68	3796.49	**3562.86**	3564.86	**3562.86**	3564.86

Table: Mean AIC copmaring six models with 50% zero proportion

Skewness	Poisson	NB	ZIP	ZINB	HP	HNB
normal	3754.84	3364.35	3011.97	3013.97	**3011.96**	3013.96
positive	3752.35	3362.43	**3012.17**	3014.17	**3012.17**	3014.17
negative	3760.90	3366.55	**3015.23**	3017.24	3015.24	3017.24

Table: Mean AIC comparing six models with 75% zero proportion

Skewness	Poisson	NB	ZIP	ZINB	HP	HNB
normal	2914.66	2169.25	**1942.16**	1944.16	1942.17	1944.17
positive	2918.74	2170.57	**1942.39**	1944.39	**1942.39**	1944.39
negative	2921.83	2171.69	**1942.89**	1944.89	**1942.89**	1944.89

Table: Mean AIC comparing six models with 90% zero proportion

Skewness	Poisson	NB	ZIP	ZINB	HP	HNB
normal	1717.94	1090.41	**982.60**	984.60	**982.60**	984.60
positive	1717.05	1081.12	**982.65**	984.65	982.66	984.66
negative	1714.56	1080.78	**982.27**	984.27	982.28	984.28

Thus, from this experiment, we could find that when the data exist the problem of zero-deflation, hurdle models will provide better model fit than zero-inflated models, yet, as zero-

proportion gets larger, the model fit of zero-inflated models get better. Another finding is that as zero-proportion gets larger, the advantages of using modified Poisson regression also gets more and more significant, thus whenever we encounter the issue of zero-inflation, we need to consider using the modified the Poisson models. What is more, in this experiment, we did not find significant advantages of using the negative binomial formulations; one reason might because the simulated data that I designed does not exist a very series issue of over-dispersion on the positive counts. The absolute over-dispersion level for skewed datasets is only 1.3, which might not large enough to illicit the advantages of using the negative binomial formulations.

Second-Simulation Findings

Model Fit

When the sample size is only 100, it is very interesting that, at different levels of zero-proportion, it is the zero-inflated negative binomial regression model that has best loglikelood(while the true distribution is zero-inflated Poision).However, since zero-inflated negative binomial regression has one extra parameter (over-dispersion parameter) than the zero-inflated model, thus as for the AIC statistics, the zero-inflated Poisson model still has best fit.(I have not finished the simulation studies yet, thus, the aic statistics table does not provide at this time).

Another finding from this table is that as the zero-proportion gets larger, the model fit for the zero-inflated and hurdle models get better, yet the model fit of Poisson and negative binomial regression get worse, which again support the results from the first experiment that the Poisson /NB models could not handle the zero-inflation issues well.

Table Mean LL for n=100, true model is ZIP

P (structrual zero)	Poisson	NB	HP	ZIP	HNB	ZINB
0.1	-536.34	-303.48	-203.10	-200.20	-203.07	**-200.16**
0.3	-1334.85	-709.57	-183.59	-180.34	-183.56	**-180.32**
0.5	-1433.84	-695.06	-145.28	-142.46	-146.65	**-143.25**
0.7	-1426.81	-734.20	-88.24	-87.00	-88.20	**-88.02**

When the sample size gets larger to 200, we could see that the zero-inflated poisson model has the best model fit, and the larger the proportion of structural zeros, the more significant advantage of the ZIP model.

Table Mean LL for n=200, true model is ZIP

P (structual zero)	Poisson	NB	HP	ZIP	HNB	ZINB
0.1	-1504.77	-559.53	-409.79	-403.27	-409.69	**-403.18**
0.3	-3153.26	-1481.55	-368.96	**-361.90**	-368.94	-361.91
0.5	-3894.77	-1980.98	-293.99	**-288.52**	-293.94	-289.38
0.7	-3457.47	-1975.96	-197.68	**-194.24**	-198.95	-195.92

The table for n=500 provides the similar results as before, and the advantages of using ZIP model become more significant than the case when n=200. It indicates that as n tures larger, the sampling distribution would get more and more closer to the true distribution.

Table Mean LL for n=500, true model is ZIP

P (structual zero)	Poisson	NB	HP	ZIP	HNB	ZINB
0.1	-3628.75	-1274.05	-1029.27	**-1014.81**	-1029.21	-1014.90
0.3	-8577.54	-2493.00	-924.24	**-905.83**	-924.20	-905.88
0.5	-10710.22	-4700.52	-754.73	**-740.23**	-754.68	-740.24
0.7	-10117.27	-5327.49	-509.25	**-502.05**	-509.17	-503.50

Abilities of Capturing Zero Observation

Another feature that we care about is the ability of predicting the zero observations, and structural zero observations (which aims to the zero-inflated models).

In the following tables, the observed zero in each datasets, together with the predicted zero observations and structural zero observations from different models are displayed for different levels of proportions of structural zero observations.

In the first table (n=100), when p=0.1, we could see that the observed zero observations has mean equals to 36, for the six models, we could see that both the zero-inflated and hurdle models has captured the zero observations very well. Here, the hurdle models have predicted the zero observations exactly equals to the observed zero, which is a result of the models' attributes, but

we could see that the zero-inflated models also has done a good job in capturing zero observations. At the same time, we could see that neither Poisson regression nor negative binomial captured enough zero observations.

Besides, we need to further focuse on the prediction of the structural zeros. Since the zero-inflated models allows the zero coming from two separate processes, thus it allows us to differientiate the different zeros. Given n=100, the expected structural zero was about 100*p. When p=0.1/0.3/0.5/0.7, the expected structural zero is 10/30/50/70, and we could see that ZIP and ZINB provides the same results. When p is small, they over-estimate the stuctural zeros, and when p is large, they under-estimate the structural zeros.

Table: Predicted zero observations for n=100, true model is ZIP

P (structual zero)	Obs	Poisson	NB	HP	ZIP	ZIP(structural zero)	HNB	ZINB	ZINB(structural zero)
10%	34	27	30	34	34	12	34	34	12
30%	48	28	37	48	48	30	48	48	30
50%	62	34	46	62	63	49	62	62	48
70%	78	37	54	78	78	70	78	78	68

When n=200, the expected structural zero would be 20/60/100/140 for each level of p, and we could see that this time, ZIP and ZINB model has captured bot the zero observations and structural zero observations very well.

Table: Predicted zero observations for n=200, true model is ZIP

P (structual zero)	Obs	Poisson	NB	HP	ZIP	ZIP(structural zero)	HNB	ZINB	ZINB(structural zero)
10%	68	54	61	68	68	22	68	68	22
30%	96	57	75	96	96	59	96	96	59
50%	125	63	89	125	125	99	125	125	99
70%	156	76	103	156	155	139	156	155	135

When n =500, the expected structural zero would be 50/150/250/350 for each level of p. This time, we found that the ZIP has better capability in predicting the structural zeros compared with the ZINB model, the reason behind this is that as n goes larger, the data gets closer to the true distribution. Again, both the zero-inflated and hurdle models have done a good job in predicting zero observations.

Table: Predicted zero observations for n=500, true model is ZIP

P (structual zero)	Obs	Poisson	NB	HP	ZIP	ZIP(structural zero)	HNB	ZINB	ZINB(structural zero)
10%	166	136	153	166	166	50	166	166	50
30%	242	155	210	242	242	151	242	242	151
50%	314	167	238	314	314	248	314	314	248
70%	388	193	259	388	388	350	388	388	346

Thus from this experiment, we could find that the zero-inflated models have very strong capabilities to predict both zero observations, and structural zero observations. Because of the model attribute, the hurdle models could always estimated the zero observations exactly match the observed zero observations, yet it assumes only one type of zero. Thus, if the research has some underlying assumptions

offf different types of zero observations, it is the zero-inflated models that would be considered when choosing models. Besides, we found that both poisson regression and negative binomial regression models have failed to capture the abundant zero observations, and the more the zero, the less the model fit. Thus, in the real research, whenever we faced the zero-inflation problem, we should always turn to the modified poisson regressions.

Reference

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: a tutorial on count regression and zero-inflated models. *Journal of Family Psychology*, *21*(4), 726.

Binkley, J. K. (2006). The effect of demographic, economic, and nutrition factors on the frequency of food away from home. *Journal of consumer Affairs*, *40*(2), 372-391.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge university press.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, 829-844.

Crowley, F., Eakins, J., & Jordan, D. (2013). Participation, expenditure and regressivity in the Irish lottery: Evidence from Irish household budget survey 2004/2005. *The Economic and Social Review*, *43*(2, Summer), 199-225.

Desouhant, E., Debouzie, D., & Menu, F. (1998). Oviposition pattern of phytophagous insects: on the importance of host population heterogeneity. *Oecologia*, *114*(3), 382-388.

Desjardins, C. D. (2013). *Evaluating the performance of two competing models of school suspension under simulation-the zero-inflated negative binomial and the negative binomial hurdle*

Famoye, F., & Singh, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*, *4*(1), 117-130.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.

Han, E., & Powell, L. M. (2013). Consumption patterns of sugar-sweetened beverages in the United States. *Journal of the Academy of Nutrition and Dietetics*, *113*(1), 43-53.

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

Lesser, L. I., Zimmerman, F. J., & Cohen, D. A. (2013). Outdoor advertising, obesity, and soda consumption: a cross-sectional study. *BMC Public Health*, *13*(1), 20.

Jaunky, V. C., & Ramchurn, B. (2014). Consumer behaviour in the scratch card market: a double-hurdle approach. *International Gambling Studies*, *14*(1), 96-114.

Jiang, Y., House, L., Tejera, C., & Percival, S. S. (2015, January). Consumption of Mushrooms: A double-hurdle Approach. In *2015 Annual Meeting, January 31-February 3, 2015, Atlanta, Georgia* (No. 196902). Southern Agricultural Economics Association.

Matheson, F. I., White, H. L., Moineddin, R., Dunn, J. R., & Glazier, R. H. (2012). Drinking in context: the influence of gender and neighbourhood deprivation on alcohol consumption. *Journal of epidemiology and community health*, *66*(6), e4-e4.

Miller, J. M. (2007). *Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation*. University of Florida.

Morland, K., Wing, S., Roux, A. D., & Poole, C. (2002). Neighborhood characteristics associated with the location of food stores and food service places. *American journal of preventive medicine*, *22*(1), 23-29.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of econometrics*, *33*(3),

341-365.
Ridout, M., Hinde, J., & DeméAtrio, C. G. (2001). A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. *Biometrics*, *57*(1), 219-223.

Shonkwiler, J. S., & Shaw, W. D. (1996). Hurdle count-data models in recreation demand analysis. *Journal of Agricultural and Resource Economics*, 210-219.

Yen, S. T., & Huang, C. L. (1996). Household demand for Finfish: a generalized double-hurdle model. *Journal of agricultural and resource economics*, 220-234.

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*(3), 275-289.