2014

A COMPARISON OF DIFFERENT METHODS OF ZERO-INFLATED DATA ANALYSIS AND ITS APPLICATION IN HEALTH SURVEYS

Si Yang
University of Rhode Island, yangsi06@gmail.com

A COMPARISON OF DIFFERENT METHODS OF ZERO-INFLATED DATA ANALYSIS AND ITS APPLICATION IN HEALTH SURVEYS

BY

SI YANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2014

MASTER OF ARTS THESIS

OF

SI YANG

APPROVED:

Thesis Committee:

Major Professor Lisa L. Harlow

Gavino Puggioni

Golleen A. Redding

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND
2014

ABSTRACT

Count data with excessive zeros and/or over-dispersion are prevalent in a wide variety of disciplines, such as public health, psychology, and environmental science. Different regression models have been proposed to deal with data with a preponderance of zero observations. These approaches include: a. transform the data to make it normal and use ordinary least-squares regression (LST); b. Poisson regression (Poisson); c. negative binomial regression (NB); d. zero-inflated Poisson regression (ZIP); e. zero-inflated negative binomial regression (ZINB); f. zero-altered Poisson regression (ZAP); and g. zero-altered negative binomial regression (ZANB). There is no clear guideline as to which one to use and it is possible that one approach is more preferable than the others under different degrees of zero-inflation and over-dispersion. This study aimed to evaluate the performance of the above seven models under different conditions of zero-inflation and over-dispersion and to examine the amount of bias and poor fit resulting from fitting various models. Simulated datasets were generated with a mixture of different proportions of zeros (20%, 40%, 60%, and 80%) and a negative binomial distribution with different dispersion parameters (10, 50, and 100). Health survey data from the Behavioral Risk Factor Surveillance System (BRFSS) study were then analyzed to further assess zero-inflated procedures and explore the relationship between physical activity and health related quality of life. Akaike Information Criterion (AIC) values and Vuong tests were used to evaluate relative quality of the regression models. Results from the simulation study showed that the ZINB and the ZANB models had smaller AIC values in all conditions of zero-inflation and over-dispersion which indicate better performance than for the other models. The LST model had the worst fit to the

data under every condition. As for the empirical study, the ZANB model was chosen as the final model and results showed that compared with highly active people, inactive people were likely to experience 1.39 more unhealthy days. Females and older people were more likely to report unhealthy days. Results also showed that estimated regression coefficients and standard errors differed across different models. There was a tendency for the worse models to have smaller standard errors and to make Type I errors. Overall, this study suggests using special zero-inflated models like ZINB or ZANB when the data have both excessive zeros and skewness in the non-zero part.

ACKNOWLEDGMENTS

First of all, I would like to thank my major professor, Dr. Lisa Harlow, who introduced me to the topic of zero-inflated data. Without her persistent support and guidance this thesis would not have been possible. I would also like to thank Dr. Gavino Puggioni for his help and assistance in programming for the simulation study. In addition, thank you to Dr. Colleen Redding for her time and helpful input. Last but not the least, I would also like to thank the CDC for collecting the BRFSS data and making it available to the public.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

FIGURE PAGE

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

In psychological, social, and public health related research, it is common that the outcomes of interest are relatively infrequent behaviors and phenomena (e.g., suicide attempts, heroin use). Data with abundant zeros are especially popular in health surveys when counting the occurrence of certain behavioral events, such as frequency of alcohol use, number of cigarettes smoked, number of hospitalizations and number of healthy days, etc. This type of data is called count data and their values are usually non-negative with a lower bound of zero and typically exhibit over-dispersion (variance much larger than mean) and/or excessive zeros.

Except for transforming the outcome to make it normal and using the general linear model, several other approaches can be taken in the context of a broader framework: generalized linear model (GLM). For example, the Poisson distribution becomes increasingly positively skewed as the mean of the response variable decreases, which reflects a common property of count data (Karazsia, Van Dulmen, 2008).Thus, a typical way of analyzing count data includes specification of a Poisson distribution with a log link, i.e. the log of a response variable is predicted by the linear combination of covariates (i.e., predictors) in a model known as Poisson regression.

Several other more rigorous approaches to analyzing count data include the zero-inflated Poisson (ZIP) model and the zero-altered Poisson model (ZAP, also called a hurdle model) that have been proposed recently to cope with an

overabundance of zeros. These last two types of models both include a binomial process (modeling zeros versus non-zeros) and a count process. The difference between the two models is how they deal with different types of zeros: while the count process of ZAP is a zero-truncated Poisson (i.e. the distribution of the response variable cannot have a value of zero), the count process of ZIP can produce zeros (Zuur, 2009). One of the assumptions of using Poisson regression is that the mean and variance of a response variable are equal. In reality, it is often the case that the variance is much larger than the mean which is called over-dispersion. Variations of negative binomial models can be used when over-dispersion exists even in the non-zero part of the distribution. While a Poisson distribution contains only a mean parameter (μ), a negative binomial distribution has an additional dispersion parameter (k) to capture the amount of over-dispersion. Thus, the zero-inflated negative binomial (ZINB) model and zero-altered negative binomial (ZANB) model were introduced to deal with both zero-inflation and over-dispersion.

To evaluate various techniques for dealing with zero-inflated count data, studies from both simulated data and empirical data have produced quite different results and model recommendations regarding which model to use. It is possible that this discrepancy resulted from not understanding the different underlying mechanism of zero-inflation and different degrees of zero-inflation and over-dispersion. It is necessary, therefore, to undertake a comprehensive examination and comparison of these methods under different conditions in order to understand how to deal with data that include too many zeros. Thus, this study proposes to answer this question and, as

an illustration, this study also applied zero-inflated models to analyze empirical data from a national health survey.

Generalized linear model (GLM) and Poisson regression

The GLM is a flexible modeling framework which allows the response variables to have a distribution form other than normal. It also allows the linear model of several covariates to be related to a response variable via arbitrary choices of link functions. Zurr et al. (2009) summarized that building a GLM consists of three steps: a) choosing a distribution for the response variable (Y); b) specifying covariates (X); and c) choosing a link function between the mean of the response variable (E(Y)) and a linear combination of the covariates (βX). Classical models such as analysis of variance (ANOVA) and ordinary least squares regression also belong to the GLM when Y is normally distributed. Y can also be specified as other distributional forms in the exponential family such as a binomial distribution, Poisson distribution, negative-binomial distribution, and gamma distribution. The link function brings together the response variable and the linear combination of the covariates. For ordinary least-squares regression, the function to estimate the expected value of Y is βX = E(Y); it is termed as an identity link. Specifying a logit link as βX = Ln (E(Y)/(1-E(Y))) is usually used for logistic regression to predict the outcome of a categorical response variable. The form of a Poisson model is as follows:

$$p(Y|X) = \frac{e^{-\mu}\mu^{Y}}{Y!} \quad y = 0, 1, 2, \ldots$$

where μ is the conditional mean count. Let $X = (X_1, \ldots, X_p)^T$ be a vector of covariates and $\beta = (\beta_1, \ldots, \beta_p)^T$ be a vector of regression parameters, where the superscript, T,

indicates a transposed vector. The logarithm of μ is assumed to be a linear combination of p covariates of the form

$$E(Y|X) = \mu = \exp(X^T\beta)$$

The conditional mean and conditional variance are equal for the Poisson regression model, that is $E(Y|X) = Var(Y|X) = \mu$. The greater the mean the greater is the variability of the data. However, over-dispersion that is present in many real-life health survey data show that a very large proportion of zeros in the count data leads to much smaller mean values than those of the variance.

Negative binomial regression model

The assumption that the variance of counts is equal to the mean also implies that the variability of the subjects sharing the same covariates values (a population has the same values for $X_1, X_2, \ldots, X_p$) is equal to the mean. If it fails to be true, the estimates of the regression coefficients can still be consistent using Poisson regression, but the standard errors can be biased. They usually tend to be too small and thus increase the rate of Type I error (false positive results). When analyzing health survey data to explore relationships between variables or make predictions, we would not expect we have measured every variable that contributes to the rates of the outcome events. There will always be residual variation in the response variables. For instance, Roebuck (2004) studied how adolescent marijuana use might relate to school attendance (estimated by number of days truant) by analyzing data from the National Household Survey on Drug Abuse. It is unlikely that adolescent marijuana users will have the same rate of being truant, specifically, there is more variation in school

attendance among marijuana users. To account for greater variation, the negative binomial model has been proposed as a generalization of the Poisson model. The negative binomial distribution is derived from a mixture of Poisson-gamma distribution in terms of mean μ and dispersion parameter k with the following form:

$$p\,(Y\,|X,\,\xi) = \left(\frac{k}{k+\mu}\right)^{r} \frac{\Gamma(k+Y)}{\Gamma(Y+1)\Gamma(k)} \left(\frac{\mu}{k+\mu}\right)^{Y}$$

where ξ is an underlying heterogeneity term which is assumed to be distributed according to a gamma *(k, k)* distribution with E (ξ) =1 and Var (ξ) = $\frac{1}{k}$. The variance of the above distribution is $\mu + \frac{\mu^2}{k}$, and hence decreasing values of *k* correspond to increasing levels of dispersion. As *k* increases towards positive infinity, a Poisson distribution is obtained. The form of a negative binomial regression model is as follows:

$$E\,(Y|X,\,\xi) = \mu\,\xi = \xi \exp\,(X^{T}\beta)$$

and this model is able to capture the over-dispersion in count data which the simple Poisson model cannot. However, the problem of excessive zeros is still not solved. For many researchers, they are also interested in finding the special meaning underlying the inflation of zeros.

Zero-inflated regression models

Statisticians have developed new approaches to model zero-inflation in count data. Lambert (1992) proposed an approach to use what is referred to as a zero-inflated Poisson (ZIP) model. In his model, two kinds of zeros are thought to exist in the data:

"structural zeros" (or true zeros) from a non-susceptible group (i.e., those that do not have the attribute or experience of interest, such as healthy people without a disease) and "random zeros" (or false zeros) for those from a susceptible group (e.g., those that have a disease in a health-based study who may falsely indicate a score of zero). The probability of being in a susceptible group π can be estimated by information from covariates with a logistic link. If an individual is from the susceptible group, his or her count is a random variable from a Poisson distribution with mean μ. The marginal distribution of the ZIP model is as follows:

$$p(Y|X) = \begin{cases} (1-\pi) + \pi e^{-\mu} & Y = 0 \\ \pi\left(\frac{e^{-\mu}\mu^{Y}}{Y!}\right) & Y > 0 \end{cases}$$

The Poisson hurdle model (i.e., ZAP) as an alternative was introduced by Mullay (1986) and modified by King (1989). It models all zeros as one part and a zero-truncated part for all non-zero observations. The main difference with ZIP is that hurdle models don't distinguish true and false zeros and all zero observations are thought to come from a non-susceptible group.

$$p(Y|X) = \begin{cases} 1-\pi & Y = 0 \\ \pi\left(\frac{e^{-\mu}\mu^{Y}}{Y!(1-e^{-\mu})}\right) & Y > 0 \end{cases}$$

Since a Poisson distribution assumes that the variance of the outcome variable equals its mean, when over-dispersion also comes from the non-zero part (i.e., the variance is much bigger than the mean even for the non-zero part), both ZIP and ZAP models can be extended to zero-inflated negative binomial (ZINB) or zero-altered negative binomial (ZANB) models to deal with zero-inflation and over-dispersion at

the same time. These types of models have become fairly popular recently and have been used to analyze number of cigarettes smoked per day (Schunck & Rogge, 2012), dental health status (Wong & Lam, 2012), depressive symptoms (Beydoun, 2012), and alcohol consumption (Atkins, 2012), etc. The major advantage of using models specially dealing with zero-inflation is that they not only reduce biases resulting from the extreme non-normality but also have the ability to model the effect on subjects' susceptibility and magnitude at the same time.

Proposed Study

For count data, depending on an outcome's mean-variance relationship and proportion of zeros, the choices for modeling its distribution range from standard Poisson and negative binomial to ZIP, and ZINB (or ZAP and ZANB). However, some researchers argue that they have seen cases where ZIP models were inadequate and ZINB also couldn't be reasonably fitted to the data (Famoye & Singh, 2006). Warton (2005) also criticized such zero-inflated models as being too routinely applied, leading to overuse. He analyzed 20 multivariate abundance datasets extracted from the ecology literature using three different approaches: least squares regression on transformed data, log-linear models (Poisson and negative binomial regression), and zero-inflated models (ZIP and ZINB), and then compared each model's goodness-of-fit. The result showed that a Gaussian (i.e., normally distributed) model (e.g., least squares regression) based on a transformed outcome fit the data surprisingly better than fitting zero-inflated count distributions. This study also suggested that negative binomial regression had the best fit and that special techniques for dealing with

excessive zeros may be unnecessary. It may be possible, however, that results reflected Type I errors based on standard errors that were too small given the model.

Based on these open questions in the field, there appears to be a conflict since there is increasing popularity of zero-inflated models, although some empirical evidence has tended to show no better fit for these models compared with the traditional least squares method conducted on transformed data. Moreover, there is much disagreement about which zero-inflated model to choose from among ZIP, ZINB, ZAP, and ZANB. In most of the zero-inflation data analysis literature, proposing an extensional zero-inflated model or comparing different models are usually motivated and illustrated by a single empirical study. These look more like case studies in which each dataset or applied situation has its particular uniqueness. It is possible that the discrepancy in the results from these studies depends on having a different proportion of zeros and different skewness in the non-zero part. It is becoming apparent that having data with excessive zeros is the norm in many situations, with or without known reasons. However, it is not clear what the proportion of zeros is, after which the data should be considered as zero-inflated, and what the underlying mechanism of abundant zeros is. Further, when researchers have collected data with abundant zeros, should zero-inflated models be used, and if so, which one should be used? These are questions that have unclear or controversial answers in the zero-inflation literature, and which are driving the proposed research. This study will use systematic methods to try to answer these questions.

Another consideration is that, the whole range of these methods hasn't been compared and tested under different conditions. The purpose of this study is to

examine the performance of different techniques dealing with zero-inflation. Both simulated data and empirical data with and without known reasons for zero-inflation are analyzed. Specifically, this study addresses the following research questions:

1. Under conditions of different zero-inflation parameters but the same over-dispersion parameter, which of the following models is superior: a) least squares regression with a transformed outcome; b) Poisson regression; c) negative binomial regression; d) ZIP; e) ZINB; f) ZAP; or g) ZANB?

2. Under conditions of different over-dispersion parameters but the same zero-inflation parameters, which of the following models is superior: a) least squares regression with a transformed outcome; b) Poisson regression; c) negative binomial regression; d) ZIP; e) ZINB; f) ZAP; or g) ZANB?

3. Finally, for the empirical data from a national health survey with a zero-inflated and over-dispersed response variable, which of the following models is superior: a) least squares regression with a transformed outcome; b) Poisson regression; c) negative binomial regression; d) ZIP; e) ZINB; f) ZAP; or g) ZANB?

CHAPTER 2

METHOD

Part 1. Simulation

Simulation Study Design A brief literature review on the frequency of various health survey outcomes showed that the percentage of zeros tends to range from 20% to 90% (Beydoun, 2012; Lin, 2012; Mahalik, 2013); thus, four conditions with varying probability of zeros (*w* in Table 1.) for the response variable were tested in the current study to reflect this range. In order to examine the effect of over-dispersion in the non-zero part, a dispersion parameter *k* with the following values: 10, 50, and 100 were pre-specified. These values represent a reasonable range of dispersion to help assess the merit of various models with varying distributions. The bigger the *k* the less dispersed the variable is and it approaches a Poisson distribution. The response variable was generated with a negative binomial distribution with a different proportion of zeros added. The simulation study was a 4 (i.e., Factor A: degree of zero-inflation) x 3 (i.e., Factor B: degree of over-dispersion) factorial design that was examined for the 7 models listed for Factor C, as shown in Table 1.

Generating Simulated datasets: To provide a reasonable prediction model to explore in this study, a count response variable Y and two different kinds of covariates, X_1 and X_2, were simulated. X_1 was assumed to be a binary variable whose values were 0 or 1 with $Pr(X_1 = 0) = Pr(X_1 = 1) = 0.5$. X_2 was set to follow a standard

normal distribution, N (0, 1). Regression coefficients β_1 and β_2 for the two covariates were set to be 0.3 and 0.5 for the population model to allow for a medium and large value, respectively. To ensure accurate results, 2000 replications (i.e., simulation size, S = 2000), each with sample size n =500, were generated. The decisions on the number of simulations and sample size were made by referring to previous simulation studies on zero-inflated data (e.g., Lambert, 1992; Min & Agresti, 2005; Williamson, 2007).

Model Selection Criteria: The model with minimum AIC (Akaike information criterion) was considered as the best model to fit the data (Bozdogan, 2000). AIC is given by:

$$AIC = -2\log L(\hat{\theta}) + 2k,$$

where $L(\hat{\theta})$ is the maximized likelihood function for the estimated model and it offers summary information on how much discrepancy exists between the model and the data, where k is the number of free parameters in the model. AIC assesses both the goodness of fit of the model and the complexity of the model. It rewards the model fit by the maximized log likelihood term, i.e., $-2\log L(\hat{\theta})$, and also prefers a relatively parsimonious model by having k as a measure of complexity.

Part 2. Empirical Data Analysis

Analyses were conducted on an existing data set to further assess different procedures. The Behavioral Risk Factor Surveillance System (BRFSS) collects information on health risk behaviors, health conditions, health care access, and use of preventive services (CDC, 2012). In this portion of the study based on actual data, the

relationship between physical activity and health related quality of life was examined after controlling for age and gender, continuous and binary covariates, respectively.

Participants

The data were obtained from the 2011 Rhode Island BRFSS, a random-digit telephone health survey of adults 18 years of age or older. Of 6533 participants involved in the survey, 38.3% were males and 61.7% were females ranging in age from 18 to 98 (M = 55.51, SD = 16.90).

Measures

Health Related Quality of Life (HRQOL): The overall number of mentally or physically unhealthy days (UNHLTH) in the last 30 days was used as an indicator of having poor HRQOL. The summary index of unhealthy days was calculated by combining the following two questions, with a logical maximum of 30 unhealthy days:

"1. Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?"

"2. Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?"

Physical Activity (PA): A set of questions was implemented in BRFSS to capture data on three key domains of physical activity: leisure-time, domestic, and

transportation. A summary score for physical activity was calculated and then was categorized into four levels, *highly active, active, insufficiently active,* and *inactive,* with higher scores indicate higher levels of physical activity.

Analysis

Participants reporting 30 physically or mentally unhealthy days during the past month were not included in the analysis. These individuals were considered as patients with long-term sickness who did not meet the inclusion criteria for this study. PA, age, gender, and their interactions with PA were entered as predictors of having poor HRQOL. Seven regression models described above were used to fit the data. In addition to using AIC values to evaluate the models, Vuong's tests were also used for model comparisons. Vuong's test is likelihood-ratio based for comparing nested, non-nested, or overlapping models in a hypothesis testing framework. The null hypothesis was that both models were equally close to the true model (Genius, 2000).

Statistical Program. R (R Core Team, 2013) was used both for data simulation and data analyses. Function rnbinom () was used to generate random negative binomial variables. Functions hurdle () and zeroinfl () from package "pscl" (Jackman, 2008) were used to fit data with zero-altered and zero-inflated models; and glm () from package "stats" was used to fit LST, Poisson, and NB models.

CHAPTER 3

RESULTS

Part 1. Results from simulation study

Figure 1.1 to 1.12 show boxplots of AIC from the seven models for every study condition. The NB, ZIP, ZINB, ZAP, and ZANB had similar AIC values which were smaller than the LST and the Poisson models under various degree of zero-inflation and over-dispersion. AIC values for the both the LST and Poisson model had a tendency to get smaller as the dispersion parameter *k* increased. However, as for the NB model, the AIC values got bigger as the dispersion parameter *k* increased.

Figures 2.1 to 2.4 show frequency plots of the simulated response variable y under 12 study conditions that varied four levels of zero-inflation and three levels of over-dispersion. The distributions show gradual changes from one condition to another: having a higher peak around zero with a bigger *w* and being more right skewed with a smaller *k*. AIC values for the four levels of zero-inflation combined with three levels of over-dispersion on the seven models are presented in Table 2. The LST model yielded the largest AIC values for each of the 12 conditions indicating the worst fit to the data.

Part 2. Results from empirical data analysis

Descriptive statistics such as means (and standard deviations) or frequencies (and percentages) for the variables of age, sex, UNHLTH and physical activity are

presented in Table 3. Participants reported an average of 3.63 unhealthy days during the past 30 days with a variance of 36.84, which was much larger than the mean; and 44.67% of the participants reported 0 unhealthy days. Figure 3 displays the frequency plot of UNHLTH which showed an extremely right skewed distribution with a spike at zero.

Seven models described above were used to fit the data. AIC values and -2log-likelihood for each model are presented in Table 4.1. The Poisson regression model had the largest AIC values, indicating a poor fit to the data. Using ordinary least-squares regression with UNHLTH log-transformed had the second largest AIC values. Of the remaining five models, the NB, ZINB, and ZANB models had smaller AICs compared with ZIP and ZAP, indicating better fit with the data. ZINB and ZANB models yielded similar AICs and are considered as the best models even after penalizing the number of parameters in the model. Since not all of the models were nested with each other, under the null hypothesis that the models were indistinguishable, Vuong tests were used to further compare the above models. OLS couldn't be compared because it has a different term for its dependent variable, i.e. it is log-transformed. The first comparison was made between the Poisson model and the NB model, with a Vuong test statistic of -42.41, and $p < 0.01$, indicating the NB model was more preferred. The more preferable model was then compared with the next model. After a series of tests and model comparisons (as shown in Table 4.2), ZANB was chosen as the best model. ZINB was the second choice with a Vuong test statistic of -1.77, and $p = 0.04$ compared to ZANB.

Regression coefficients and standard errors were estimated and presented in Table 5.1 and 5.2 for each of the seven models when applied to the BRFSS dataset. Standard errors estimated from different models were quite different. There was a tendency for the worse models to have smaller standard errors, thus to make Type I errors. For instance, although estimates from the Poisson model were similar to those from the NB model, their standard errors were much smaller, thus yielding significant results for most of the regressors. It was the same when comparing ZIP versus ZINB and ZAP versus ZANB.

With PA, gender, age, PA*gender, and PA*age predicting both the count model and zero-inflation model, Table 5.2 shows parameter estimates from the ZANB model (the final model). Participants in the highly active group and males were used as reference groups. After controlling for age, gender, and their interaction terms with PA, compared with highly active people, inactive people were likely to experience 1.39 (=exp(0.326), $p < 0.001$) more unhealthy days. This trend can also be seen in Figure 4, where both inactive males and females had higher means of UNHLTH than other groups of participants. Gender (odds ratio = 1.27, $p < 0.05$) and age (odds ratio = 0.99, $p < 0.001$) were the only results to be significant predictors for those who experienced 0 unhealthy days versus those who experienced more than 0 unhealthy days. Females and older people were more likely to report unhealthy days, although it should be pointed out that the odds ratio for age was not very meaningful in size, even if significant.

R code for the simulation study, and R and SAS code for the BRFSS data analyses are provided after the Tables and Figures at the end of the Appendices.

CHAPTER 4

DISCUSSION

This study evaluated seven regression models under various conditions of zero-inflation and over-dispersion by analyzing simulated datasets and an empirical dataset. Results from both studies suggested that when the data include excessive zeros and are over-dispersed, models using negative binomial regression or zero inflated models (i.e. ZIP, ZINB, ZAP, and ZANB) perform better than Poisson regression and ordinary least-squares regression with the response variable transformed (LST). Fitting data with the smallest proportion of zeros and less dispersion (i.e., $w = 0.2$ and $k = 100$) in the simulation study, the Poisson regression model had the least discrepancy on its AIC values compared with other preferred models. However, the LST model still performed much worse than the other models.

The poor fit from the LST might be that the log-transformation still fails to correct the non-normality and to address the inflation of zeros. Another drawback of using transformation is that the regression coefficients are harder to interpret. The Poisson distribution is the probability model usually assumed for count data, however, zero-inflated count data usually tend to have much bigger variance than the mean which violates its assumption that the mean equals the variance. In both cases, when failing to address the problem of zero-inflation and over-dispersion, standard errors of the estimates tended to be deflated or under estimated (Hilbe, 2011). Furthermore, if inappropriately choosing the LST or the Poisson model, there is greater tendency to

make Type I errors, i.e. a variable may appear to be a significant predictor when it is in fact, not significant. Estimated regression coefficients from Table 5.1 demonstrate this kind of bias.

Results from this study support using special zero inflated models for zero-inflated data. When over-dispersion also exists even in the non-zero part of the data, a negative binomial regression instead of the regular Poisson regression should be used. The use of zero inflated models can be justified on both substantive and statistical grounds. Substantively, zero inflated models have the ability to identify the factors that have significant effects on the probability that the participant is from the non-susceptible group by means of a binary regression model; and the magnitude of the counts given that the participant is from the susceptible group by means of a Poisson regression or negative binomial regression. Factors or explanatory variables do not need to be the same for the binomial model and the count model. Although the NB model can also effectively offer accurate estimation under some degrees of zero-inflation and over-dispersion, it cannot provide information about possible mechanisms underlying the zero-inflation. Statistically, zero inflated models provide more accurate estimates as shown by both the simulation results and empirical data analysis results.

Zero-inflated models and zero-altered models produced very similar results from both simulated data and the BRFSS data. The decision when choosing between these two should rely on the nature of the research questions. The biggest difference between them is that zero-inflated models distinguish between structural zeros (true zeros) and random zeros (false zeros), while zero-altered models do not. In public

health and medicine studies, zero-inflated models may be conceptualized as allowing zeros to arise from at-risk (susceptible) and not-at-risk (non-susceptible) populations. In contrast, we may conceptualize zero-altered models as having zeros only from an at-risk population (Rose et al., 2006). For instance, when answering a survey question that asks the number of drinks someone had during the past month, some people report 0 drinks because they are abstainers and they never drink. However, for people who are regular drinkers, they might also report 0 drinks if they did not drink during that month. As mentioned earlier, we call these zero responses random zeros. It is more appropriate to use ZIP and ZINB in these kind of situations when the study design has a greater chance of having random zeros. One the other hand, however, zero-inflated and zero-altered modeling framework should be equivalent when the primary purpose of the study is to make predictions as both of them tend to yield similar model fit.

Another interest of the study was to explore the relationship between health related quality of life (HRQoL) and physical activity (PA). Many research studies have shown that PA helps to improve overall health and fitness, and reduce risk of health conditions including diabetes, coronary heart disease, stroke, and cancers (CDC, 2014). Despite the well-known benefits of exercise, according to the CDC, less than half of American adults meet the recommended level of PA. HRQoL describes both the physical and mental well-being of an individual. It is an important concept in health research and can help to inform decisions on the prevention and treatment of diseases. The present study examined the relationship between PA and HRQoL after controlling for relevant demographic characteristics within the context of a large representative health survey from Rhode Island. Results showed that participants

reporting higher levels of PA tended to report fewer unhealthy days. Specifically, compared with participants in the highly active group, those who seldom reported any physical activity were likely to experience 1.30 more unhealthy days. Females and older people were also more likely to report unhealthy days versus 0 unhealthy day compared to males and younger people. These findings offer a better understanding that health-related lifestyle behaviors, such as being more physically active, can improve HRQoL and might help to inform policy makers to provide more intervention programs for the general population.

There were also some limitations of the study. First, findings from the simulation study were only based on a limited number of conditions. Simulation results offer a general picture as to which model is more appropriate, however, more conditions will need to be examined to get a more accurate relationship between the model selection and different levels of zero-inflation and over-dispersion. In practice, it is still suggested to fit a range of possible models and choose one that has relatively better fit and can also answer the research questions. Second, explanatory variables for the zero versus non-zero model and the count model were set to be the same. The most attractive advantage of using zero-inflated models is that they allow researchers to have different predictors for two parts of the models, which usually can be justified theoretically. Since the data were collected via a telephone survey, various response biases and non-response biases would occur. For instance, participants consist mostly of older people with an average age of 55.51 years, thus the sample was not sufficiently random. Third, the cross-sectional nature of the data was another limitation of the study. Since these data were cross-sectional, no temporal order can be

determined, so it is possible that those with higher health-related quality of life (HRQoL) reported more physical activity (PA). Future longitudinal designs will be needed to tease out temporal relationships. Only age and gender were controlled for in the empirical data analysis. It is possible that other unmeasured factors, such as disease states and seasonality, could be potential confounding variables of the relationship between PA and HRQoL. Future longitudinal analyses would help to improve our understanding of these relationships and increase the predictive power of the study, in addition to what model is used to examine the data. Finally, the UNHLTH ranges from 0 to 29 days, which follows a zero-inflated negative binomial distribution truncated at 29. Creel, et al. (1990) suggest that accounting for truncation of the response variable provides a more accurate coefficient estimates, regardless of the choice of the statistical model. Although a truncated model was not used in this study, it might be of interest in future studies.

APPENDICES

Tables

Table 1. Simulation design factors

Factor A:	**Factor B:**	**Factor C:**
w	***k***	**Models (Tested on each of the 4x3 conditions in A & B)**
0.20	10	Least squares regression with transformed outcome (LST)
0.40	50	Poisson regression model (Poisson)
0.60	100	Negative binomial regression model (NB)
0.80		Zero-inflated Poisson model (ZIP)
		Zero -inflated negative binomial model (ZINB)
		Zero -altered Poisson model (ZAP)
		Zero -altered negative binomial model (ZANB)

Table 2. AIC for 12 conditions on 7 models

	LST	**Poisson**	**NB**	**ZIP**	**ZINB**	**ZAP**	**ZANB**
w=0.8 & k=10	1225.31	712.17	569.91	560.26	561.37	560.54	561.69
w=0.8 & k=50	1200.87	708.54	576.89	564.71	566.30	565.03	566.66
w=0.8 & k=100	1195.82	706.47	576.73	564.48	566.13	564.85	566.53
w=0.6 & k=10	1519.14	1059.51	925.92	911.20	911.78	911.62	912.22
w=0.6 & k=50	1490.53	1049.04	931.93	912.88	914.37	913.35	914.87
w=0.6 & k=100	1484.49	1046.79	932.43	912.91	914.52	913.40	915.03
w=0.4 & k=10	1648.39	1267.42	1180.08	1166.58	1166.61	1166.95	1167.01
w=0.4 & k=50	1612.40	1250.50	1181.38	1162.73	1164.18	1163.23	1164.67
w=0.4 & k=100	1606.84	1248.51	1182.21	1163.12	1164.68	1163.61	1165.18
w=0.2 & k=10	1693.61	1385.43	1349.19	1343.77	1343.18	1343.93	1343.50
w=0.2 & k=50	1648.16	1361.98	1340.18	1330.75	1332.04	1331.06	1332.43
w=0.2 & k=100	1640.83	1358.68	1338.96	1329.08	1330.55	1329.37	1330.90

Table 3. Descriptive statistics for independent and dependent variables (n = 5670)

Variable	**Mean (SD) or Frequency (%)**
Age	55.03 (16.87) years
Sex	
Male	2126 (38.7%)
Female	3362 (61.3%)
# Unhealthy Days	3.63 (6.07) days
Physical Activity	
Highly Active	1659 (32.5%)
Active	1059 (20.8%)
Insufficiently Active	1059 (20.8%)
Inactive	1323 (25.9%)

Table 4.1. Model fit comparison from the BRFSS data

	LST	**Poisson**	**NB**	**ZIP**	**ZINB**	**ZAP**	**ZANB**
AIC	33170.83	47932.45	21447.22	27814.26	21060.95	27814.26	21060.06
-2log-likelihood	33194.83	47908.45	21421.22	27766.26	21010.95	27766.26	21010.06
(df)	(13)	(12)	(13)	(24)	(25)	(24)	(25)

Table 4.2. Vuong non-nested tests results

Model Comparison	Vuong Test Statistic (*p*)	Preferable Model
Poisson vs. NB	-41.42 (< 0.01)	**NB**
NB vs. ZIP	22.30 (< 0.01)	**NB**
NB vs. ZINB	-12.16 (< 0.01)	**ZINB**
ZINB vs. ZAP	25.35 (< 0.01)	**ZINB**
ZINB vs. ZANB	-1.77 (0.04)	**ZANB**

Table 5.1. Estimated regression coefficients (and standard errors)

Regressor	**LST**	**Poisson**	**NB**
Intercept	0.713***	0.987***	0.983***
	(0.040)	(0.023)	(0.080)
PA_active	0.032	0.097*	0.116
	(0.068)	(0.038)	(0.134)
PA_insufficiently active	-0.004	0.021	0.027
	(0.068)	(0.039)	(0.133)
PA_inactive	0.162**	0.360***	0.365**
	(0.062)	(0.033)	(0.122)
SEX_female	0.117**	0.173***	0.178
	(0.053)	(0.029)	(0.104)
AGE	-0.007***	-0.010***	-0.010**
	(0.002)	(0.000)	(0.003)
PA_active*SEX_female	0.049	-0.002	-0.025
	(0.086)	(0.046)	(0.169)
PA_insufficiently	0.158	0.231***	0.225
active*SEX_female	(0.085)	(0.046)	(0.168)
PA_inactive*SEX_female	0.110	0.089*	0.083
	(0.080)	(0.040)	(0.157)
PA_active*AGE	0.001	0.004**	0.005
	(0.003)	(0.001)	(0.005)
PA_insufficiently	0.005	0.009***	0.009
active*AGE	(0.003)	(0.001)	(0.005)
PA_inactive*AGE	0.007**	0.012***	0.012**
	(0.002)	(0.001)	(0.005)

Table 5.2. Estimated regression coefficients (and standard errors) continue

Regressor	**ZIP**	**ZINB**	**ZAP**	**ZANB**
	Count Model			
Intercept	1.903***	1.754***	1.903***	1.753***
	(0.023)	(0.065)	(0.023)	(0.065)
PA_active	0.047	0.051	0.047	0.055
	(0.038)	(0.105)	(0.038)	(0.106)
PA_insufficiently active	0.000	-0.001	0.000	-0.001
	(0.039)	(0.106)	(0.039)	(0.106)
PA_inactive	0.281***	0.325***	0.281***	0.326***
	(0.033)	(0.095)	(0.033)	(0.095)
SEX_female	0.039	0.046	0.039	0.046
	(0.030)	(0.082)	(0.030)	(0.082)
AGE	-0.002*	-0.002	-0.002*	-0.002
	(0.001)	(0.002)	(0.001)	(0.002)
PA_active*SEX_female	-0.044	-0.047	-0.044	0.051
	(0.047)	(0.129)	(0.046)	(0.129)
PA_insufficiently	0.123**	0.143	0.123**	0.142
active*SEX_female	(0.046)	(0.149)	(0.046)	(0.129)
PA_inactive*SEX_female	0.015	0.007	0.015	0.005
	(0.041)	(0.119)	(0.041)	(0.120)
PA_active*AGE	0.002	0.002	0.002	0.007
	(0.001)	(0.004)	(0.001)	(0.003)
PA_insufficiently	0.005***	0.005	0.005***	0.053
active*AGE	(0.001)	(0.004)	(0.001)	(0.004)
PA_inactive*AGE	0.006***	0.007*	0.006***	0.007*
	(0.001)	(0.003)	(0.001)	(0.003)

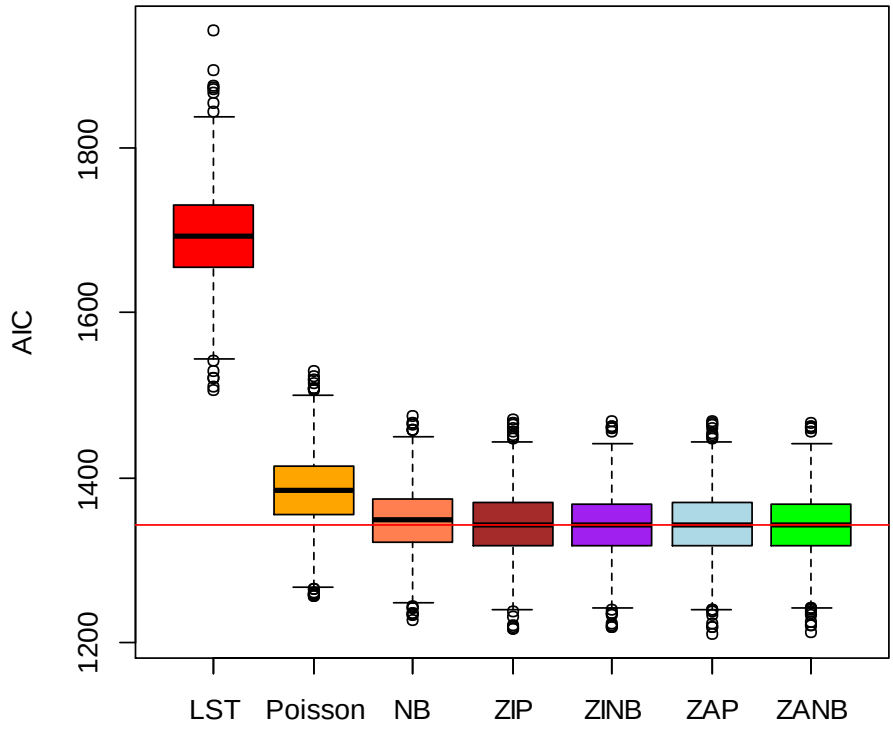
Regressor	**ZIP**	**ZINB**	**ZAP**	**ZANB**
		Zero-inflation Model		
Intercept	0.393***	0.127	-0.395***	-0.395***
	(0.078)	(0.092)	(0.078)	(0.078)
PA_active	-0.074	-0.074	0.075	0.075
	(0.131)	(0.151)	(0.131)	(0.131)
PA_insufficiently active	-0.018	-0.019	0.018	0.018
	(0.131)	(0.150)	(0.130)	(0.130)
PA_inactive	-0.123	-0.060	0.125	0.125
	(0.120)	(0.135)	(0.120)	(0.120)
SEX_female	-0.126*	-0.256*	0.236*	0.236*
	(0.102)	(0.118)	(0.102)	(0.102)
AGE	0.015***	0.017***	-0.015***	-0.015***
	(0.003)	(0.004)	(0.003)	(0.003)
PA_active*SEX_female	-0.103	-0.129	0.102	0.102
	(0.165)	(0.193)	(0.165)	(0.165)
PA_insufficiently	-0.226	-0.223	0.228	0.228
active*SEX_female	(0.164)	(0.192)	(0.164)	(0.164)
PA_inactive*SEX_female	-0.170	-0.184	0.170	0.170
	(0.154)	(0.175)	(0.154)	(0.154)
PA_active*AGE	-0.002	-0.001	0.002	0.002
	(0.005)	(0.006)	(0.005)	(0.005)
PA_insufficiently	-0.008	-0.007	0.008	0.008
active*AGE	(0.005)	(0.006)	(0.005)	(0.005)
PA_inactive*AGE	-0.010*	-0.010*	0.010*	0.010*
	(0.004)	(0.005)	(0.004)	(0.004)

Notes: “Male” was the reference group for the variable sex and “highly active” was the reference group for the variable physical activity.

Significance levels: ‘***’ 0.001, ‘**’ 0.01, ‘*’ 0.05

Figures

Figure 1.1. Boxplot of AIC from seven models (w = 0.2 and k = 10)



Note: A reference line was added to Figures 1.1 to 1.12 by using the ZIP model's mean AIC values.

Figure 1.2. Boxplot of AIC from seven models (w = 0.2 and k = 50)

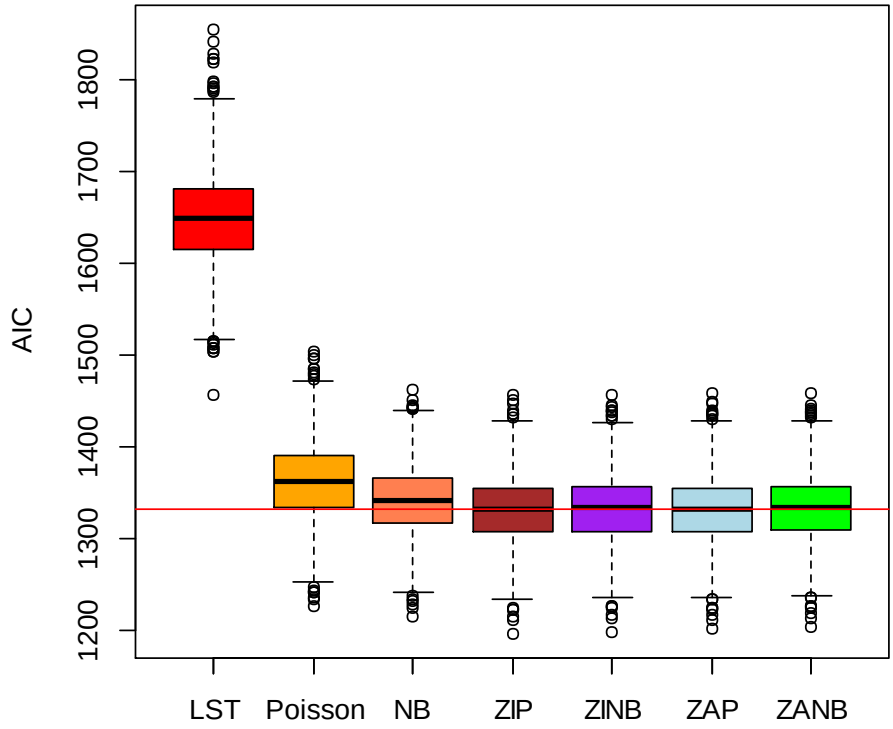


Figure 1.3. Boxplot of AIC from seven models (w = 0.2 and k = 100)

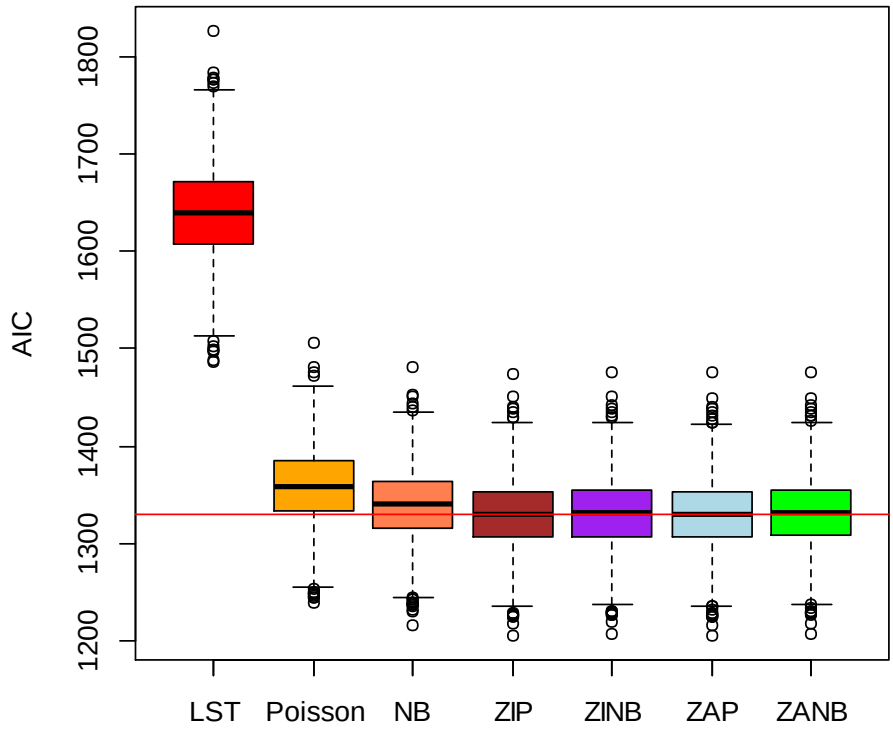


Figure 1.4. Boxplot of AIC from seven models (w = 0.4 and k = 10)

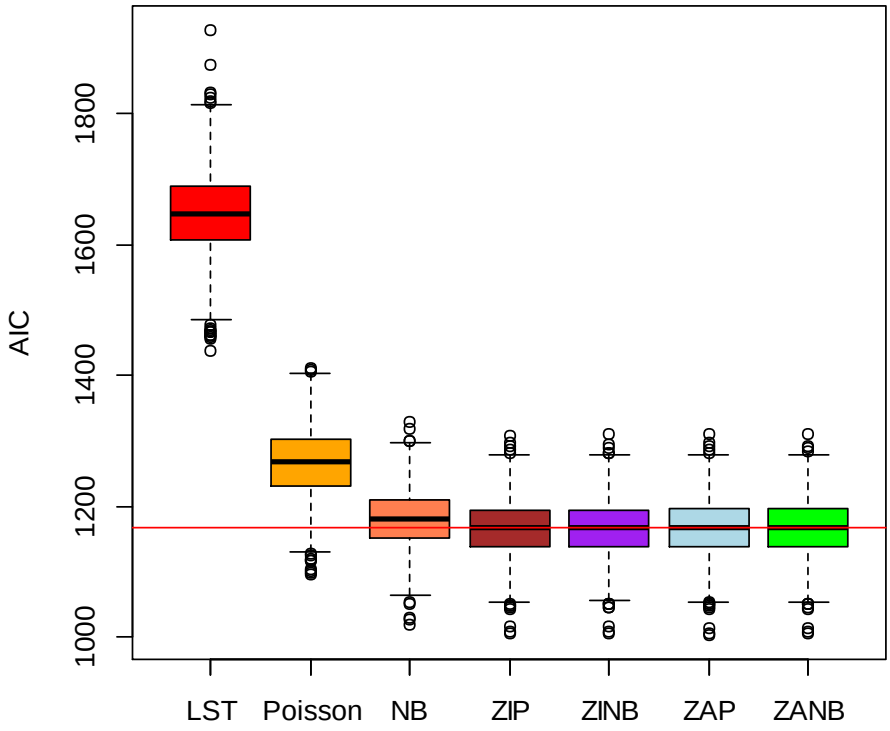


Figure 1.5. Boxplot of AIC from seven models (w = 0.4 and k = 50)

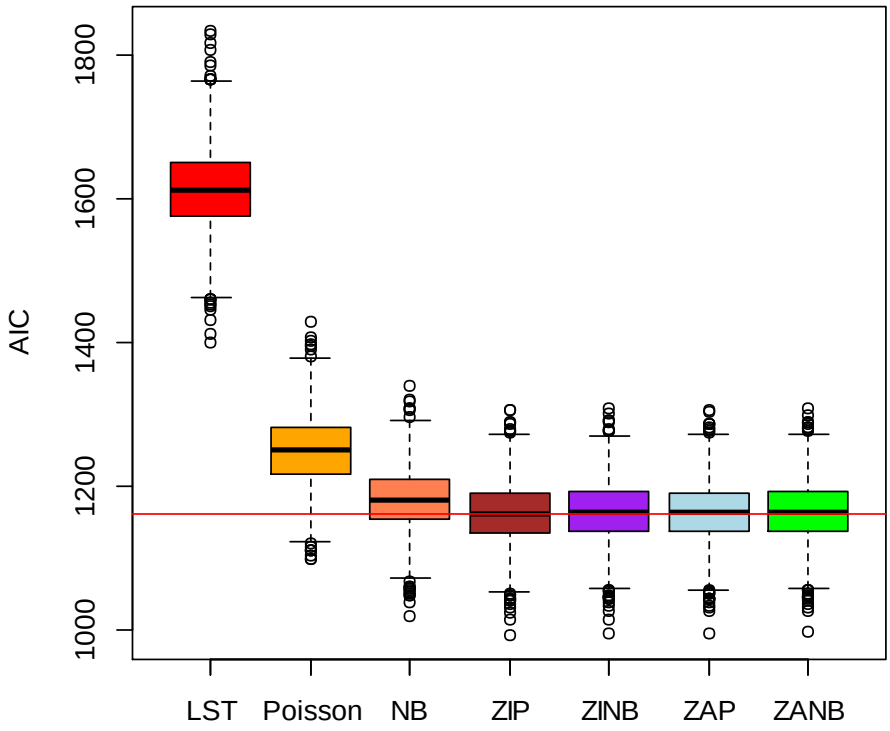


Figure 1.6. Boxplot of AIC from seven models (w = 0.4 and k = 100)

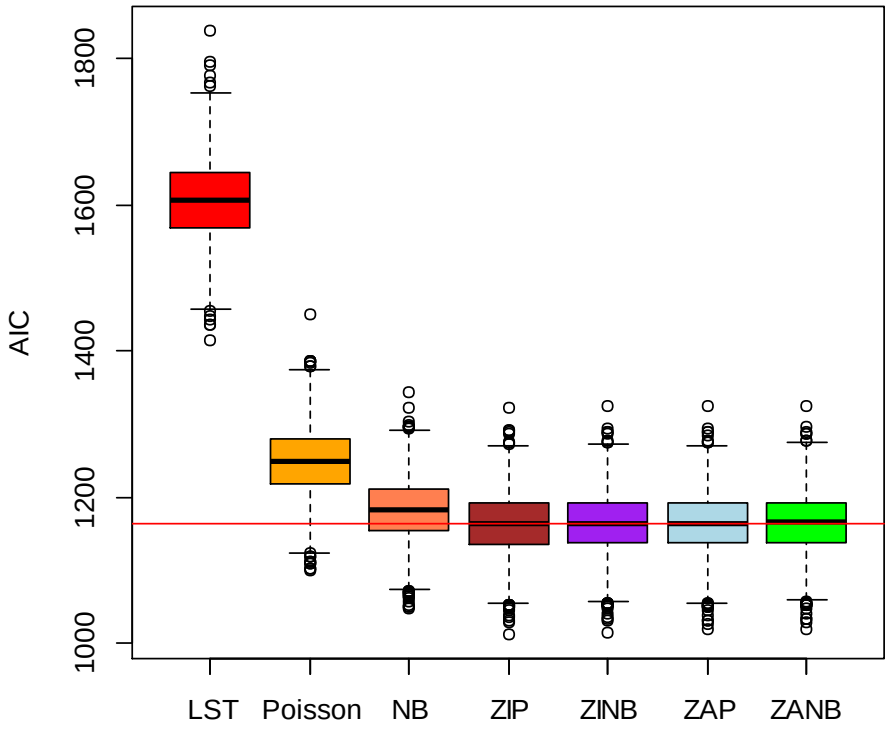


Figure 1.7. Boxplot of AIC from seven models (w = 0.6 and k = 10)

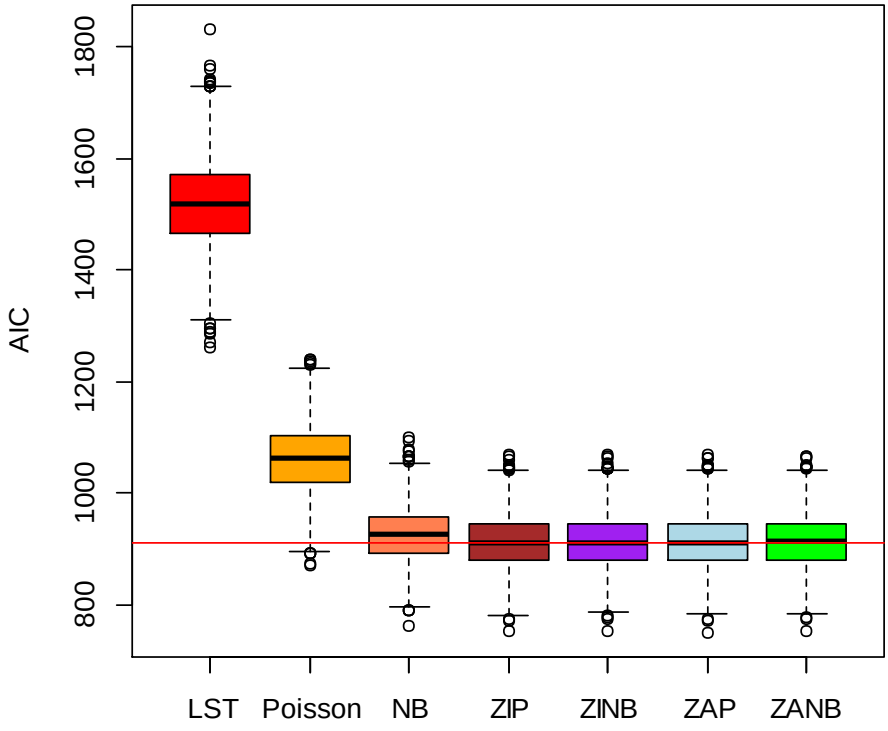


Figure 1.8. Boxplot of AIC from seven models (w = 0.6 and k = 50)

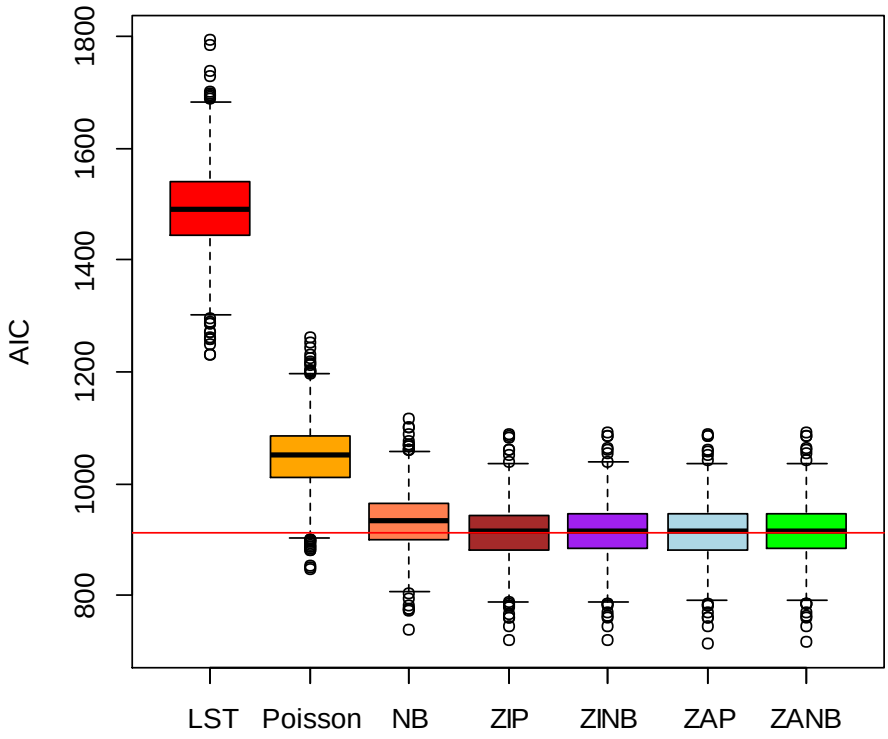


Figure 1.9. Boxplot of AIC from seven models (w = 0.6 and k = 100)

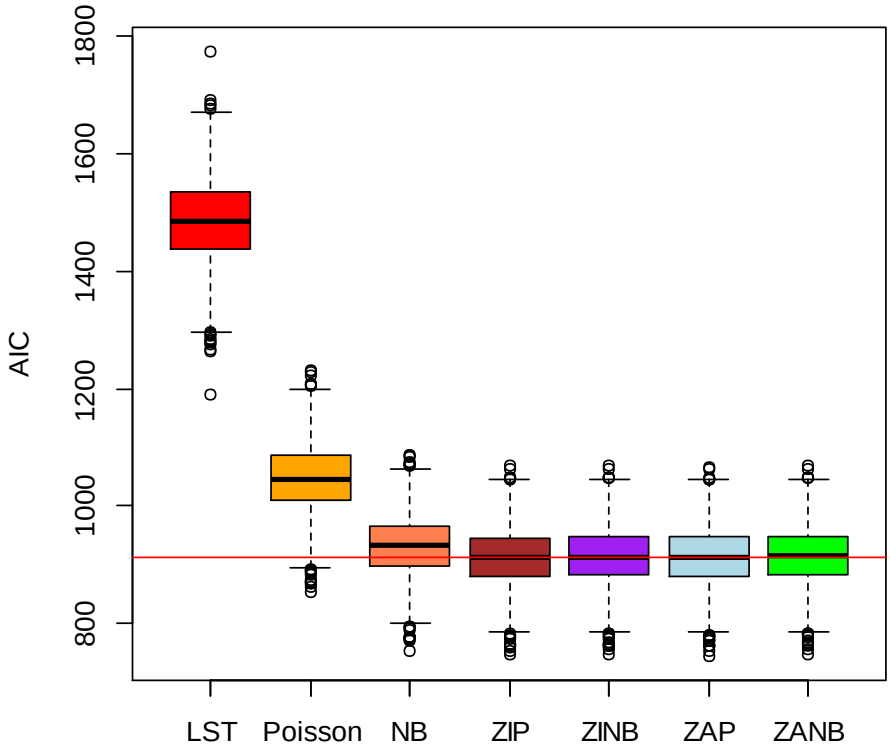


Figure 1.10. Boxplot of AIC from seven models (w = 0.8 and k = 10)

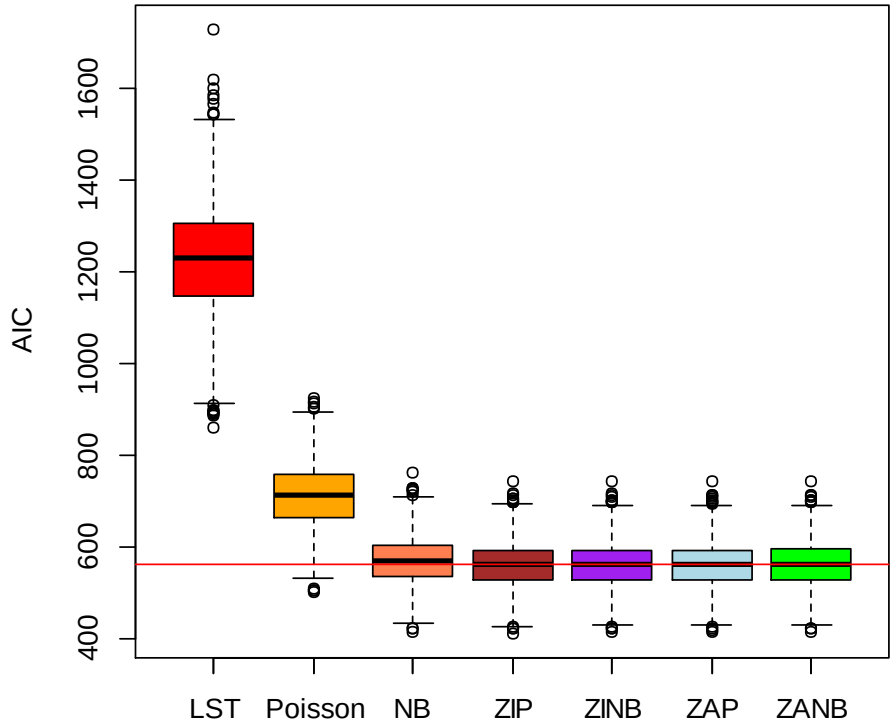


Figure 1.11. Boxplot of AIC from seven models (w = 0.8 and k = 50)

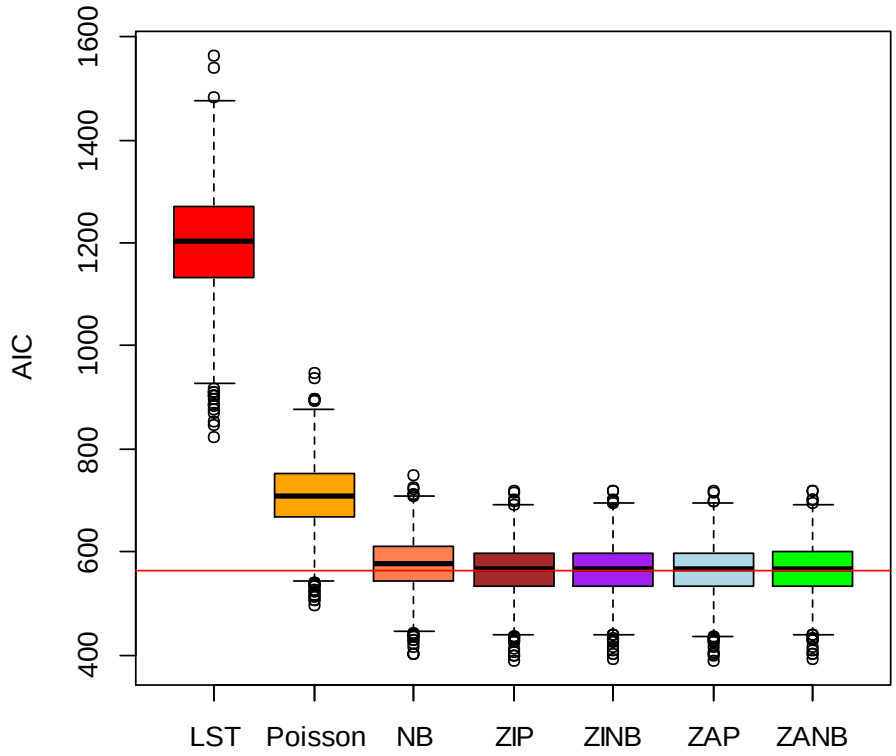


Figure 1.12. Boxplot of AIC from seven models (w = 0.8 and k = 100)

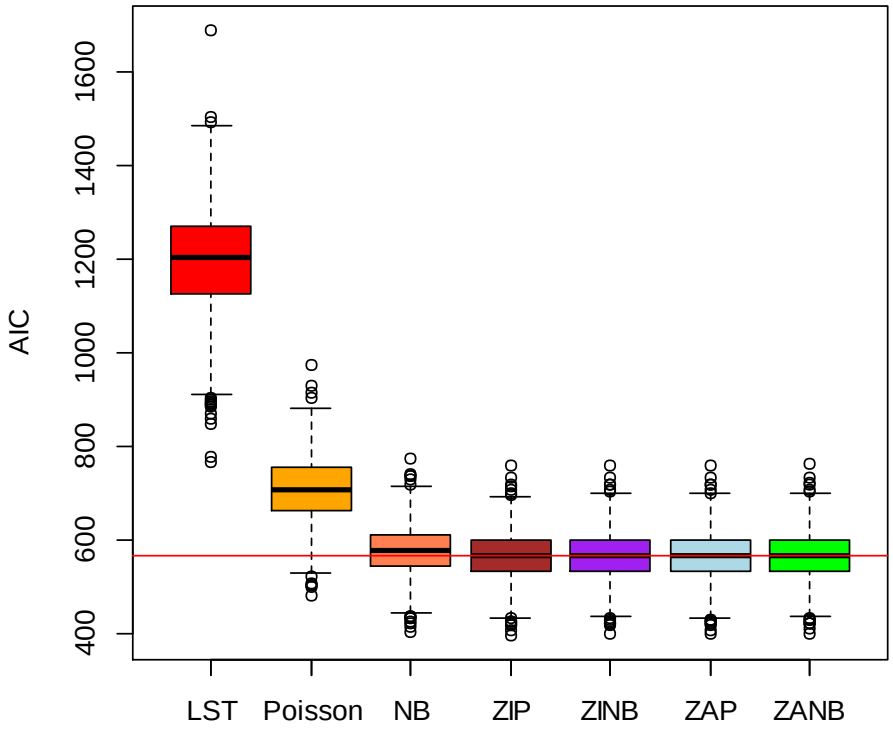


Figure 2.1. Frequency plot of simulated response variable y (w=0.2)

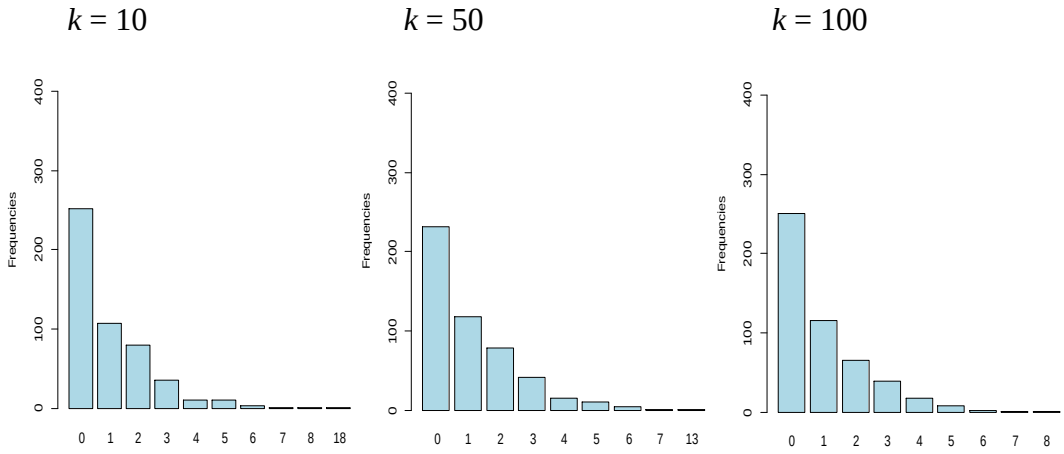


Figure 2.2. Frequency plot of simulated response variable y (w=0.4)

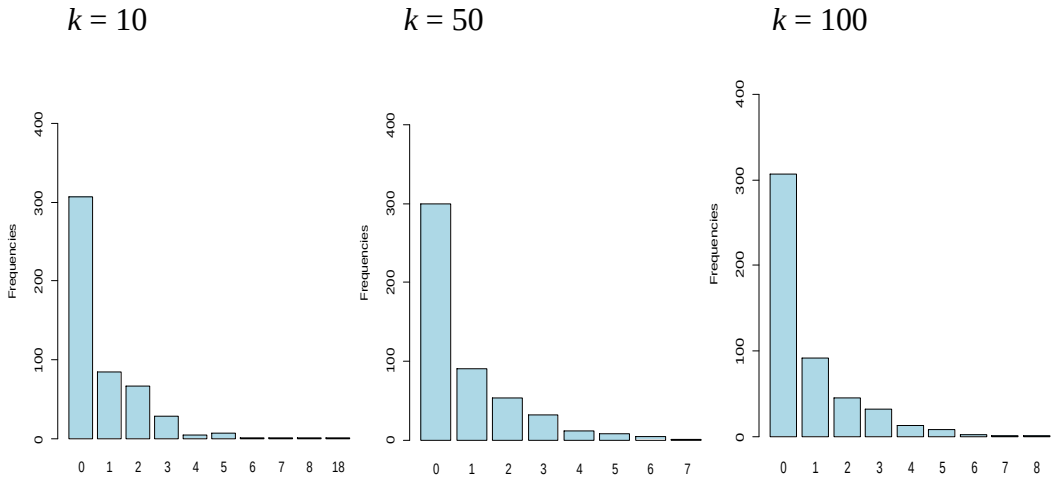


Figure 2.3. Frequency plot of simulated response variable y (w=0.6)

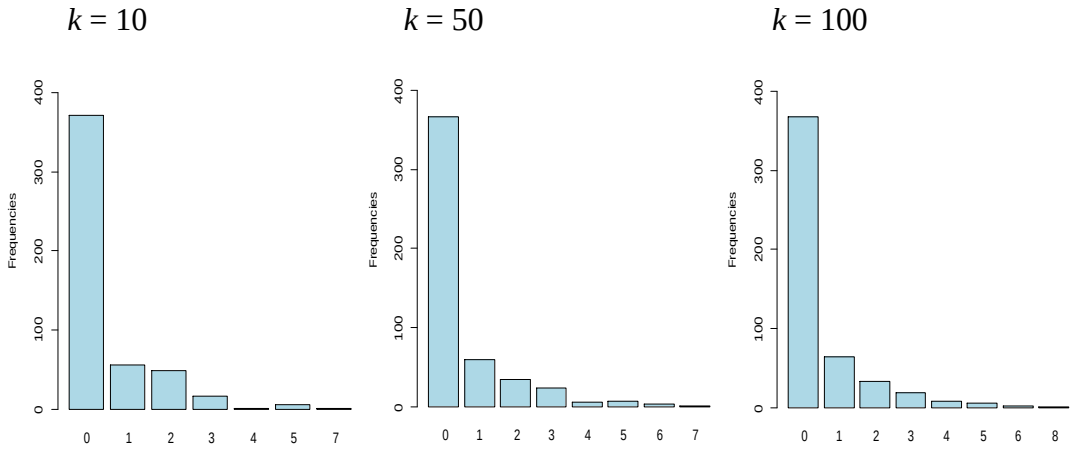


Figure 2.4. Frequency plot of simulated response variable y (w=0.8)

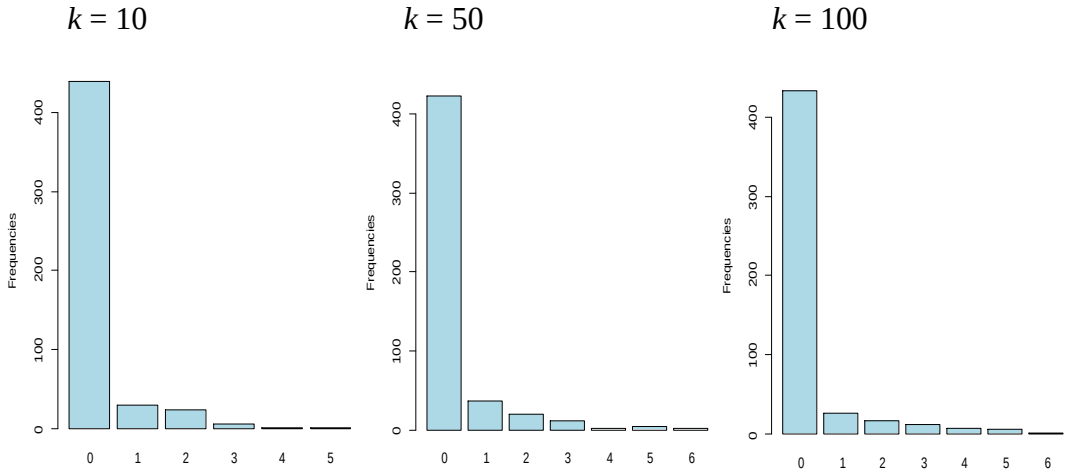


Figure 3. Frequency plot of the response variable UNHLTH from BRFSS data

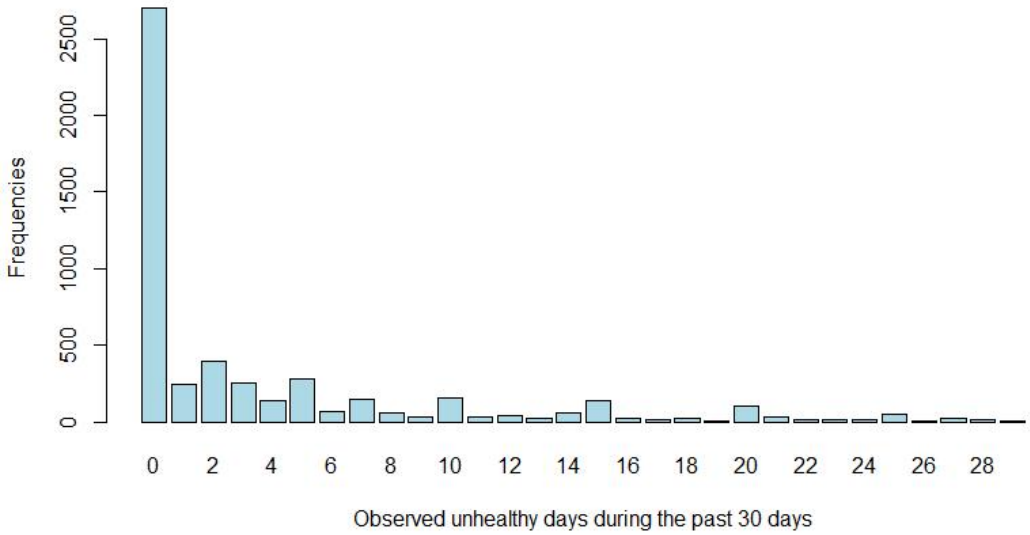
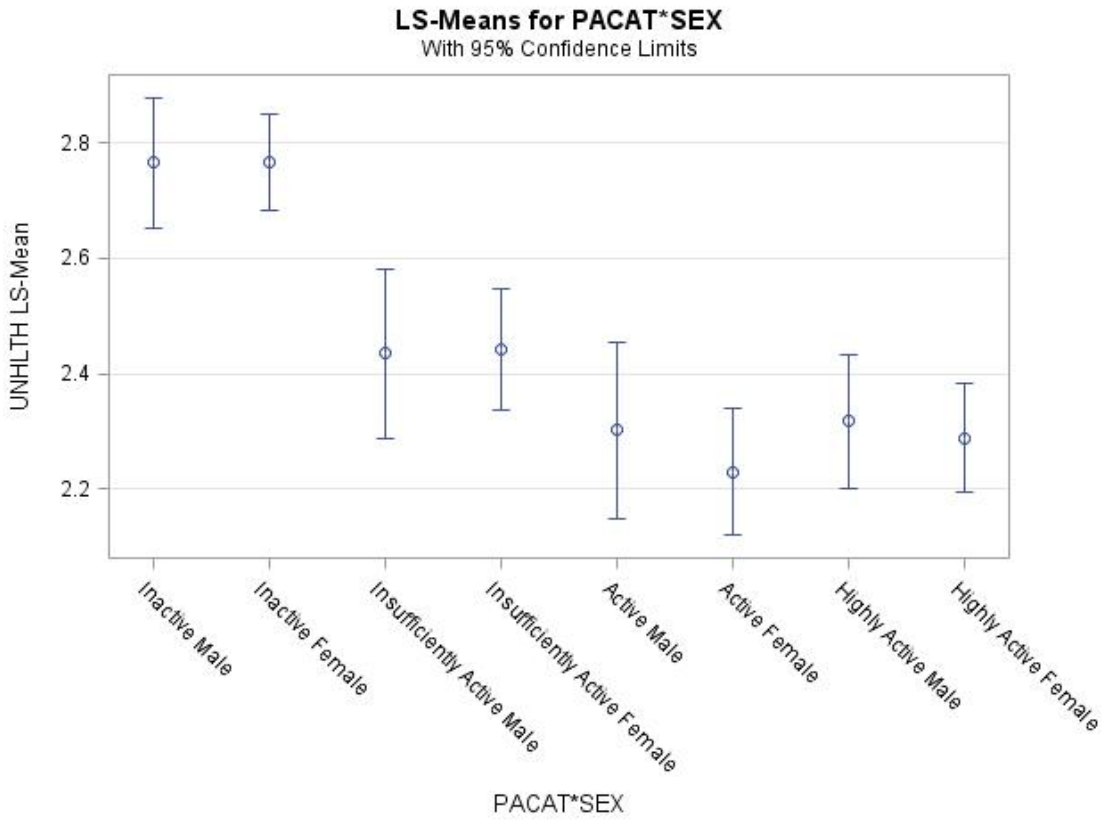


Figure 4. Least-squared Means of UNHLTH by PA and Gender with 95% Confidence Limits



R code and SAS code.

1. R code for the simulation study.

```
library(psych)
library(pscl)
library(gmodels)


set.seed(2)
S<- 2000 # Simulation size
n<- 500 # Sample size
b0<- 0 # intercept
b1<-.3 # slope for x1, categorical variable
b2<-.5 # slope for x2, continuous variable
pb<-txtProgressBar()

##Create containers for saving results
y<- matrix(0,S,1)
x1<- matrix(0,S,1)
x2<- matrix(0,S,1)

loglik1<-matrix(0,S,1)
AIC1<-matrix(0,S,1)
loglik2<-matrix(0,S,1)
AIC2<-matrix(0,S,1)
loglik3<-matrix(0,S,1)
AIC3<-matrix(0,S,1)
```

```
loglik4<-matrix(0,S,1)
AIC4<-matrix(0,S,1)
loglik5<-matrix(0,S,1)
AIC5<-matrix(0,S,1)
loglik6<-matrix(0,S,1)
AIC6<-matrix(0,S,1)
loglik7<-matrix(0,S,1)
AIC7<-matrix(0,S,1)

for (i in 1:S) {
 setTxtProgressBar(pb, i/S)
 ##generate the categorical covariate
 x1 <- sample( c(0,1), n, replace=TRUE, prob=c(0.5, 0.5) )
 ##generate the continuous covariate
 x2 <- rnorm(n,0,1)

 ##form the expected value (i.e.mean) of the response variable
 mean<-exp(b0 + b1*x1 + b2*x2)
 ##generate response variable with zero-inflated negative binomial distribution
 #Probability of zeros(w): .2, .4, .6, .8
 #Over-dispersion parameter(size k): 5, 10, 100
 #16 conditions
 w<-runif(n)
 y<- (w>.8)*(rnbinom(n,mu=mean,size=100))
 ty<- log(y+1) #log transformed y
```

```
  ##save the ith generated dataset

  data_i<-data.frame(y=y,ty=ty,x1=x1,x2=x2)


  ##fit the data into different models


  model1<- glm(ty ~ x1 + x2, data=data_i) ##Ordinary linear regression with DV log
transformed

  model2<- glm(y ~ x1 + x2, family = poisson, data=data_i) ##Poisson regression

  model3<- glm.nb(y ~ x1 + x2, data=data_i) ##negative binomial regression

  model4<- zeroinfl(y ~ x1 + x2, data=data_i) ##zero-inflated Poisson regression

  model5<- zeroinfl(y ~ x1 + x2, dist="negbin", data=data_i) ##zero-inflated negative
binomial regression

  model6<- hurdle(y ~ x1 + x2, data=data_i) ## Poisson hurdle

  model7<- hurdle(y ~ x1 + x2, dist="negbin", data=data_i) ## negative binomial
hurdle


  ###Model 1: OLS estimation with log transformed outcome

  ##store the loglikelihood and AIC values

  yf1<-exp(model1$fitted.values)-1

  res1<-y-yf1

  loglik1[i,]<-sum(dnorm(y,yf1,sqrt(sum(res1^2)/(n-2-1)),log=T))

  AIC1[i,]<- -2*loglik1[i,]-2*2


  ###Model 2: Poisson regression

  ##store the loglikelihood and AIC values
```

```
loglik2[i,]<-logLik(model2)
AIC2[i,] <-  AIC(model2)

###Model 3: Negative binomial regression
##store the loglikelihood and AIC values
loglik3[i,]<-logLik(model3)
AIC3[i,] <-  AIC(model3)

###Model 4: Zero-inflated Poisson regression
##store the loglikelihood and AIC values
loglik4[i,]<-logLik(model4)
AIC4[i,] <-  AIC(model4)

###Model 5: Zero-inflated negative binomial regression
##store the loglikelihood and AIC values
loglik5[i,]<-logLik(model5)
AIC5[i,] <-  AIC(model5)

###Model 6: Hurdle Poisson regression
##store the loglikelihood and AIC values
loglik6[i,]<-logLik(model6)
AIC6[i,] <-  AIC(model6)

###Model 7: Hurdle negative binomial regression
##store the loglikelihood and AIC values
```

```
 loglik7[i,]<-logLik(model7)
 AIC7[i,] <-  AIC(model7)
}

close(pb)

result.AIC<- cbind(AIC1, AIC2, AIC3, AIC4, AIC5, AIC6, AIC7)
result.loglik<- cbind(loglik1, loglik2, loglik3, loglik4, loglik5, loglik6, loglik7)

#### Descriptive results for AIC and Log-likelihood
describe(result.AIC)
describe(-2*result.loglik)
####
barplot(table(y), ylab="Frequencies", ylim=c(0,450), cex.lab=.9, col="lightblue")
x11()
####Boxplot of AIC from seven models
boxplot(result.AIC[,1], result.AIC[,2], result.AIC[,3], result.AIC[,4],
      result.AIC[,5], result.AIC[,6], result.AIC[,7],border=par("fg"),
      col=c("red","orange","coral","brown", "purple","light blue","green"),
      ylab="AIC")
axis(side=1, 1:7, c("LST","Poisson","NB","ZIP","ZINB","ZAP","ZANB"))
abline (h=mean (result.AIC[,4]), col="red") # add reference line
```

2. R code for analyzing the BRFSS data

```
BRFSS<-read.csv("C:/Users/siyang/Desktop/thesis for psych/BRFSS.csv")

# Select relevant variables for the analysis

myvar<-c("UNHLTH", "PACAT", "SEX", "AGE")

BRFSS<-BRFSS[myvar]

ZERO<-BRFSS[!(BRFSS$UNHLTH==30),] # Delete UNHLTH = 30

ZERO$PACAT[ZERO$PACAT == 9]<-NA # 9's are missing values

ZERO$tUNHLTH <- log(ZERO$UNHLTH+1) #log transform DV

ZERO$PA<- factor(ZERO$PACAT, labels=c("higly active","active","insufficiently
active","inactive"),exclude=NA)

ZERO$SEX<-factor(ZERO$SEX, labels=c("male","female"))

ZERO$CAGE<-(ZERO$AGE-55.03) # center continuous variable AGE

ZERO<-ZERO[complete.cases(ZERO),] #delete obs with missing values


# Basic descriptive stats

describe(ZERO)

CrossTable(ZERO$PA)

CrossTable(ZERO$SEX)

barplot(table(ZERO$UNHLTH), ylab="Frequencies",

      xlab="Observed unhealthy days during the past 30 days",

      main="Figure 1 Frequency Plot of the Response Variable UNHLTH from
BRFSS Data",

      font.main=3, col="lightblue")


hist(ZERO$UNHLTH)

corr.test(ZERO)
```

```
NONZERO<-subset(ZERO, UNHLTH>0)

describe(NONZERO) # Get mean and variance for the non-zero part


##fitting different regression models

model1<- glm(tUNHLTH ~ PA + SEX + CAGE + PA*SEX + PA*CAGE,
data=ZERO) ##ordinary linear regression with transformed outcome

model2<- glm(UNHLTH ~ PA + SEX + CAGE + PA*SEX + PA*CAGE,
data=ZERO, family = poisson) ##Poisson

model3<- glm.nb(UNHLTH ~ PA + SEX + CAGE + PA*SEX + PA*CAGE,
data=ZERO) ##negative binomial

model4<- zeroinfl(UNHLTH ~ PA + SEX + CAGE + PA*SEX + PA*CAGE,
data=ZERO) ##zero-inflated Poisson

model5<- zeroinfl(UNHLTH ~ PA + SEX + CAGE + PA*SEX + PA*CAGE,
data=ZERO, dist="negbin") ##zero-inflated negative binomial

model6<- hurdle(UNHLTH ~ PA + SEX + CAGE + PA*SEX + PA*CAGE,
data=ZERO) ## Poisson hurdle

model7<- hurdle(UNHLTH ~ PA + SEX + CAGE + PA*SEX + PA*CAGE,
data=ZERO, dist="negbin") ## negative binomial hurdle

#get model summaries

summary(model1)

summary(model2)

summary(model3)

summary(model4)

summary(model5)

summary(model6)

summary(model7)


#vuong tests to compare model fit
```

```
vuong(model2, model3)

vuong(model3, model4)

vuong(model3, model5)

vuong(model5, model6)

vuong(model5, model7)


#get AIC and loglikehood to compare model fit, the smaller the better fit

#AIC for model 1

PRED_UNHLTH<-exp(model1$fitted.values)-1

res1<-ZERO$UNHLTH-PRED_UNHLTH

loglik1<-sum(dnorm(ZERO$UNHLTH,PRED_UNHLTH,sqrt(sum(res1^2)/(5080-2-
1)),log=T)) # n=5080 complete cases

-2*loglik1

AIC1<- -2*loglik1-2*12 # number of parameters in model 1 is 12

AIC(model2)

AIC(model3)

AIC(model4)

AIC(model5)

AIC(model6)

AIC(model7)

-2*logLik(model1)

-2*logLik(model2)

-2*logLik(model3)

-2*logLik(model4)

-2*logLik(model5)
```

```
-2*logLik(model6)
-2*logLik(model7)
#get the number of parameters in each model
length(coef(model1))
length(coef(model2))
length(coef(model3))
length(coef(model4))
length(coef(model5))
length(coef(model6))
length(coef(model7))

# Odds ratio for significant predictors from the count model
exp(.325268)#PA inactive
exp(.006979) # PA inactive: AGE

# Odds ratio for significant predictors from the zero-inflation model
exp(-.255475) # SEX female
exp(.0165554) # AGE
exp(-.010288) # PA inactive:AGE

#get 95% CI of the odds ratio
exp(confint(model5))
```

3. SAS code for analyzing the BRFSS data

```
Proc import out = WORK.BRFSS
            datafile= 'C:\2013fall\STA 542\project\BRFSS.csv'
            dbms=csv replace;
Run;

proc contents data=brfss;
run;

/*Descriptive Stats*/
proc freq data=brfss;
table sex pacat unhlth;
run;

proc means data=brfss;
var age unhlth;
run;

proc univariate data = brfss noprint;
     histogram unhlth / midpoints = 0 to 30 by 1 vscale =
count;
run;


proc format;
value sexf 0='Female' 1='Male';
value pacatf 1='Highly Active' 2='Active' 3='Insufficiently
Active' 4='Inactive';
run;

proc sort data=brfss;
     by descending pacat;
run;

/*Model 1: Normal Regression with log link*/
proc genmod data=brfss order=data plots=all;
class pacat sex;
model unhlth = pacat|sex pacat|age/dist= normal link=log;
format sex sexf. pacat pacatf.;
run;


/*Model 2: Poisson Regression*/
proc genmod data=brfss order=data plots=all;
class pacat sex;
model unhlth = pacat|sex pacat|age/dist= poisson link=log;
format sex sexf. pacat pacatf.;
run;
```

```
/*Model 3: Negative Binomial Regression*/
proc genmod data=brfss order=data plots=all;
class pacat sex;
model unhlth = pacat|sex pacat|age/dist= negbin link=log;
format sex sexf. pacat pacatf.;
run;

/*Model 4: Zero-Inflated Poisson Regression*/
proc genmod data=brfss order=data plots=all;
class pacat sex;
model unhlth = pacat|sex pacat|age/dist= zip;
zeromodel pacat|sex pacat|age/ link=logit;
format sex sexf. pacat pacatf.;
run;


/*Model 5: Zero-Inflated Negative Binomial Regression*/
proc genmod data=brfss order=data plots=all;
class pacat sex;
model unhlth = pacat|sex pacat|age/dist= zinb;
zeromodel pacat|sex pacat|age/ link=logit;
lsmeans pacat*sex/cl;/*get sex by PA interaction plot*/
format sex sexf. pacat pacatf.;
run;


/*Model 6: Poisson Hurdle Regression*/
proc fmm data=brfss order=data;
class pacat sex;
model unhlth = pacat|sex pacat|age/ dist=TruncPoisson;
model       +            / dist=Constant;
run;

/*Model 7: Negative Binomial Hurdle Regression*/
proc fmm data=brfss order=data;
class pacat sex;
model unhlth = pacat|sex pacat|age/ dist=Truncnegbin;
model       +            / dist=Constant;
run;
```

BIBLIOGRAPHY

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., & Neighbors, C. (2012). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*. *A*dvance online publication. doi: 10.1037/a0029508.

Beydoun, M.A., Beydoun, H.A., Boueiz, A., Shroff, M.R., & Zonderman, A.B.(2012). Antioxidant status and its association with elevated depressive symptoms among US adults: National Health and Nutrition Examination Surveys 2005-6. *British Journal of Nutrition. A*dvance online publication. doi: 10.1017/S0007114512003467.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology, 44(1),* 62-91.

Centers for Disease Control and Prevention. (2012). *Behavioral risk factor surveillance system*. Retrieved from http://www.cdc.gov/brfss/

Centers for Disease Control and Prevention (CDC). (2014). *Facts about Physical Activity*. Retrieved from http://www.cdc.gov/physicalactivity/data/facts.html

Creel, M. D., & Loomis, J. B. (1990). Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California. *American Journal of Agricultural Economics, 72(2),* 434-441.

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

Fletcher, D., MacKenzie, D., & Villouta, E. (2005). Modeling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics, 12,* 45-54.

Gao, S. Y., Mokhtarian, P.L., Johnston, R.A. (2008). Nonnormality of data in structural equation models. *Transportation Research Record, 2082*(1), 116-124.

Genius, M., & Strazzera, E. (2002). A note about model selection and tests for non-nested contingent valuation models. *Economics Letters, 74(3),* 363-370.

Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge University Press.

Jackman S. (2008). *PSCL: classes and methods for R developed in the political science computational laboratory, Stanford University*. Department of Political Science, Stanford University, Stanford, California. R package version 1.04.4, URL http://CRAN.R-project.org/package=pscl.

Karazsia, B.T., van Dulmen, M.H. (2008) Regression models for count data: illustrations using longitudinal predictors of childhood injury. *Journal of Pediatric Psychology, 33*(10), 1076-1084.

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

Lin, T. H., &Tsai, M.H. (2012) Modeling health survey data with excessive zero and *K* responses. *Statistics in Medicine. A*dvance online publication. doi: 10.1002/sim.5650.

Liu H, &Power DA. (2007). Growth curve models for zero-inflated count data: an application to smoking behavior. *Structural Equation Modeling*, *14*, 247–79.

Long, J. (1997). *Regression models for categorical and limited dependent variables.* CA: Thousand Oaks, Sage.

Mahalik, J. R., Levine Coley, R., McPherran Lombardi, C., Doyle Lynch, A., Markowitz, A. J., & Jaffee, S. R. (2013). Changes in health risk behaviors for males and females from early adolescence through early adulthood. *Health Psychology.* Advance online publication. doi: 10.1037/a0031658

Mullay, J. (1986). Specifications and testing of some modified count data model. *Journal of Econometrics, 33*, 341–365.

Myers, J.L., Well, A.D., Lorch, R.E. (2002) *Research design and statistical analysis*, 3rd Ed. Routledge.

Molas, M., Lesaffre, E. (2010). Hurdle models for multilevel zero-inflated data via H-likelihood. *Statistics in Medicine*, *29*, 3294-3310.

Min Y., Agresti A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modeling*, *5*, 1–19.

Ma, R., Hasan, M.T., Sneddon G. (2009). Modeling heterogeneity in clustered count data with extra zeros using compound Poisson random effect. *Statistics in Medicine*, *28*, 2356–2369

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ridout, M.S., Hinde, J.P. and Demetrio, C.G.B. (2001).A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, *57*, 219-223.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics, 16(4),* 463-481.

Rosner, B. (2011), *Fundamentals of biostatistics*, 7th Edition. MA: Boston, Brooks/Cole.

Schunck, R., Rogge, B.G. (2012). No causal effect of unemployment on smoking? A German panel study. *International Journal of Public Health, 57*(6), 867-874.

Tu, W. (2002). Zero-Inflated Data. *Encyclopedia of environmetrics. Vol. 4* (eds. EI-Shaarawi, A. H., Piegorsch, W.W.), Wiley, 2387-2391.

Van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics, 51,* 738–743.

Vives, J., Losilla, J. M., & Rodrigo, M. F. (2006). Count data in psychological applied research. *Psychological Reports, 98*(3), 821–835.

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.

Williamson, J. M., Lin, H., Lyles, R. H., & Hightower, A. W. (2007). Power calculations for ZIP and ZINB models. *Journal of Data Science, 5*(4), 519-534.

Wong, K.Y., & Lam, K.F. (2012). Modeling zero-inflated count data using a covariate-dependent random effect model. *Statistics in Medicine*. *A*dvance online publication. doi: 10.1002/sim.5626

Zorn, C. (1996). Evaluating zero-inflated and Hurdle Poisson specifications, *Midwest Political Science Association, 18(20),* 1–16.

Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A., & Smith, G.M. (2009). *Mixed effects models and extensions in ecology with R*. NY: New York, Springer.