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

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,