THE IMPACT OF MISSPECIFICATION OF NUISANCE PARAMETERS ON TEST FOR HOMOGENEITY IN ZERO-INFLATED POISSON MODEL: A SIMULATION STUDY

by

Siyu Gao

B.S., Huazhong University of Science and Technology, China, 2012

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2014

Approved by:

Major Professor
Wei-Wen Hsu

Copyright

Abstract

The zero-inflated Poisson (ZIP) model consists of a Poisson model and a degenerate distribution at zero. Under this model, zero counts are generated from two sources, representing a heterogeneity in the population. In practice, it is often interested to evaluate this heterogeneity is consistent with the observed data or not. Most of the existing methodologies to examine this heterogeneity are often assuming that the Poisson mean is a function of nuisance parameters which are simply the coefficients associated with covariates. However, these nuisance parameters can be misspecified when performing these methodologies. As a result, the validity and the power of the test may be affected. Such impact of misspecification has not been discussed in the literature. This report primarily focuses on investigating the impact of misspecification on the performance of score test for homogeneity in ZIP models. Through an intensive simulation study, we find that: 1) under misspecification, the limiting distribution of the score test statistic under the null no longer follows a chi-squared distribution. A parametric bootstrap methodology is suggested to use to find the true null limiting distribution of the score test statistic; 2) the power of the test decreases as the number of covariates in the Poisson mean increases. The test with a constant Poisson mean has the highest power, even compared to the test with a well-specified mean. At last, simulation results are applied to the Wuhan Inpatient Care Insurance data which contain excess zeros.

Key words: zero-inflated Poisson model, score test, misspecification, nuisance parameter

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my sincere appreciation to Dr. Wei-Wen Hsu, my major professor, for his great help, invaluable suggestions, insightful comments and patient guidance, and all the time he dedicated for this report.

I would also like to thank Dr. Weixing Song and Dr. Weixin Yao, for their willingness to serve on my committee, and making comments and sharing ideas.

Especially, I am heartily thankful to Dr. Weixing Song, my program advisor, for his warm encouragement and patient guidance during my course study, which enabled me better understand the major of statistics and improve my performance.

I would like to show my gratitude to China Life Insurance Company Limited for providing the 2012-2013 Inpatient Care Insurance data. Without data the application study in this report would have not been possible.

My special thanks go to all my instructors, for their wonderful classes and rich knowledge during my graduate study. They let me enjoy the world of statistics. Data, I'm lov'in it! Zeal for data triggers my strong interest in Statistics. This inspires me devote my dedication and unsparing effort to do this report at Kansas State University.

Finally, I am willing to thank everyone in the Department of Statistics at Kansas State University for their kindness and help.

Dedication

This is dedicated to my girlfriend, parents and grandparents, for their deep love, warm encouragement and forever support.

Chapter 1

Introduction

In China, with an increasing number of people purchasing the health insurance products, especially the Inpatient Care Insurance, the insurance companies have begun to pay more attention to the number of claims. Poisson regression model is the most popular model to analyze these count data. However, claim data usually contain excess zeros and the standard Poisson regression model may fit inadequately. Instead, the zero-inflated Poisson (ZIP) regression can be used to handle excess zeros, see Lambert (1992). The ZIP model consists of a Poisson model and a degenerate distribution at zero. In this model, both the Poisson mean and the mixing weight can depend on covariates, where the mixing weight is a probability of an excess zero. This is a very attractive feature because the number of claims is often assumed to be affected by some potential factors, for example, age, gender, occupation and living habits.

Under the ZIP model, zero counts are generated from the Poisson component and the degenerated distribution at zero. Thus a heterogeneity is present in the population. In practice, it is often interested to evaluate this heterogeneity is consistent with the observed data or not. In the literature, there are several tests can be used to evaluate this heterogeneity. For example, a score test proposed by van den Broek (1995), can be used to examine heterogeneity in ZIP models by testing whether the mixing weight equals zero or not, where

he assumed a constant mixing weight under the alternative. Jansakul and Hinde (2002) extended his test to allow that the mixing weight can depend on covariates via an identity link function under the alternative. However, the identity link function in Jansakul's methodology may need to be constrained when fit the model and it is rarely used. Todem and Hsu (2012) developed a score test for homogeneity in a more general way via a novel transformation where mixing weight also depends on covariates under the alternative. Most of the existing methodologies for evaluating heterogeneity in ZIP models are often assuming that the Poisson mean is a function of nuisance parameters which are the coefficients associated with covariates. However, these nuisance parameters can be misspecified when performing these methodologies. As a result, the validity and the power of the test may be affected. Many papers have mentioned this type of issue under several settings. For example, Godfrey (1988) pointed out that the misspecification may affect the Lagrange multiplier test in regression models. Bera and Yoon (1993) showed that the score test is not robust when nuisance parameter is locally misspecified (which assumes that the misspecification occurs from the local data generating process). Liang and Self (1996) also indicated that the nuisance parameter may be misspecified in likelihood functions, which could affect both the validity and the power of the likelihood ratio test. Aerts et al.(1999) mentioned the impact of likelihood misspecification on the robustness of lack-of-fit tests. These authors mentioned that the misspecification could affect both the validity and the power of the test. For tests of homogeneity in ZIP models, the misspecification of nuisance parameters may also have an impact on the test and it is unclear in the literature.

In this report, we focus on how the misspecification of nuisance parameters— in our case is the misspecification of the Poisson mean— could affect the power of the homogeneity test for ZIP models. The homogeneity test in this study is simply testing whether the mixing weight equals to zero or not. We use the score test in this study rather than the likelihood ratio test and the Wald test, because the score test doesn't require the model under the alternative hypothesis to be estimated. However, this doesn't mean that the score test is

better. These three tests are asymptotically equivalent, for example see Molenberghs and Verbeke (2007), and ones may choose one of these three tests based on the difficulty of computations in their particular problems. Through an intensive simulation study, we find that under the misspecification of the Poisson mean, the limiting distribution of the score test statistic under the null no longer follows a χ^2 distribution. By using a parametric bootstrap methodology to find the true null limiting distribution, the test with a constant Poisson mean outperforms the other tests with the means that depends on corvariates. The power decreases as the number of covariates in the Poisson mean increases.

The application is implemented to study the Wuhan Inpatient Care Insurance data which we mentioned at the beginning of this chapter. We want to investigate whether a standard Poisson regression fit the data adequately and what factors affect the number of claims significantly. The test results show that the data are not in flavor of Poisson regression model. Instead, we use ZIP model to fit the data and we find 6 significant factors affecting the number of claims: age, marital status, monthly income, BMI, smoking and drinking habit.

The layout of this report is as follows. In chapter 2, we present a brief review of the ZIP regression model and a general score test. In chapter 3, we discuss three types of misspecification of nuisance parameters and present the parametric bootstrap methodology. An intensive simulation study and an application of the Wuhan Inpatient Care Insurance data are present in chapter 4. Some discussions and conclusions are provided in the last chapter.

Chapter 2

Models and Test Statistics

2.1 Zero-Inflated Poisson distribution

Consider independent descrete random variables Y_i with a zero-inflated Poisson distribution, the probability mass function is given by

$$Pr(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)e^{-\lambda_i}, & y_i = 0, \\ (1-\omega_i)\frac{e^{-\lambda_i}\lambda_i{}^{y_i}}{y_i!}, & y_i = 1, 2, ... \end{cases} \tag{2.1}$$

where ω_i is the mixing weight and $0 \leqslant \omega_i \leqslant 1$. We denote this by $Y_i \sim ZIP(\lambda_i, \omega_i)$.

The ZIP model can be regarded as a simple two-component mixture model with a Poisson(λ_i) component and a degenerate component putting all its mass at zero with a probability ω_i. It is obvious that the ZIP model reduces to the standard Poisson model when $\omega_i = 0$. For positive values of ω_i we have zero-inflation, however, it is possible for $\omega_i < 0$ under the marginal ZIP model and to still obtain a valid probability distribution which leads to the zero-deflated Poisson model. An extended mixture model in which ω_i is not constrained to be a non-negative is commonly referred to as a zero-modified model. Details of these are given in Dietz and Böhning (2000).

For observations $y_1, ..., y_n$, the log-likelihood function of the ZIP model is given by

$$\begin{aligned} l &= l(\boldsymbol{\lambda}, \boldsymbol{\omega}; \mathbf{y}) \\ &= \sum_{i=0}^{n} \Big[I_{y_i=0} \log[\omega_i + (1-\omega_i)e^{-\lambda_i}] + I_{y_i>0}[\log(1-\omega_i) - \lambda_i + y_i \log \lambda_i - \log(y_i!)] \Big], \end{aligned} \tag{2.2}$$

where $\mathbf{y} = (y_1, y_2, ...y_n)^T$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ...\lambda_n)^T$, $\boldsymbol{\omega} = (\omega_1, \omega_2, ...\omega_n)^T$ and $I_{(.)}$ is the indicator function for the specified event, i.e. equals to 1 if the event is true and 0 otherwise;

To apply the zero-inflated Poisson model in practical modeling situations, Lambert (1992) suggested to use the following joint models for $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$

$$\log(\boldsymbol{\lambda}) = X\boldsymbol{\beta} \qquad and \qquad \log\Big(\frac{\boldsymbol{\omega}}{1-\boldsymbol{\omega}}\Big) = G\boldsymbol{\gamma}, \tag{2.3}$$

where X and G are covariate matrices and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ are $p \times 1$ and $q \times 1$ vectors of unknown parameters. It may be also useful to apply a identity link function for $\boldsymbol{\omega}$,

$$\log(\boldsymbol{\lambda}) = X\boldsymbol{\beta} \qquad and \qquad \boldsymbol{\omega} = G\boldsymbol{\gamma}, \tag{2.4}$$

Maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be obtained by standard approaches for mixture models: the EM-algorithm or Newton-Raphson. However, some disadvantages with the identity link are that the model fitting may need to be constrained and it is rarely used. To overcome such limitations, Todem and Hsu (2012) developed a score test where ω depends on covariates in a more general way through a novel transmation, for details see Todem and Hsu (2012).

2.2 Test statistics for homogeneity

The homogeneity test conducted in our report is just testing the mixing weight $\boldsymbol{\omega}$

$$H_0 : \boldsymbol{\omega} = 0 \qquad vs \qquad H_1 : \boldsymbol{\omega} > 0,$$

There are 3 tests can be used for testing the homogeneity in the ZIP model: the likelihood ratio test, the Wald test and score test. However, the score test statistic has an advantage over the likelihood ratio test and the Wald test, for it only requires the parameter estimates under the null hypothesis. Because of this attracting advantage, we use score test rather than the likelihood ratio test and the Wald test. Here we briefly present a general form of the score test statistic proposed by Jansakul and Hinde (2002). More details about the likelihood ratio test and the Wald test see Appendix A.

2.2.1 A score test for homogeneity in ZIP models

We use the score test proposed by van den Broek (1995), which is a special case of Jansakul's general score test by assuming a constant model for $\boldsymbol{\omega}$—taking G in (2.4) to be a $n \times 1$ matrix of 1's. In this report, we use Jansakul's expressions to introduce the score test statistic. In our study, we assume that $\boldsymbol{\omega} = \gamma_0$, then testing $\boldsymbol{\omega} = 0$ is equivalent to testing $\gamma_0 = 0$ in the complex model.

Since the score test only requires the maximum likelihood estimates of the parameters under the null hypothesis, we just need to fit the standard Poisson model. Based on the log-likelihood function given in (2.2) and the general model equations (2.4), the score test statistic under the null hypothesis is (details of the derivation of the score test statistic are in Appendix B):

$$S_{\boldsymbol{\omega}} = S_{\boldsymbol{\gamma}}{}^T(\hat{\boldsymbol{\beta}}_0, 0) C^{-1} S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0),$$

where $\hat{\boldsymbol{\beta}}_0$ is the maximum likelihood estimate under the Poisson model and

$$S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0) = G^T \Big[\frac{I_{y_i=0} - e^{-\hat{\boldsymbol{\lambda}}_0}}{e^{-\hat{\boldsymbol{\lambda}}_0}} \Big],$$

$$C = J_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0) - J_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0)^T J_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0, 0)^{-1} J_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0),$$

with

$$J_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0, 0) = X^T diag(\hat{\boldsymbol{\lambda}}_0) X,$$

$$J_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0) = G^T diag(\frac{1 - e^{-\hat{\boldsymbol{\lambda}}_0}}{e^{-\hat{\boldsymbol{\lambda}}_0}}) G,$$

and

$$J_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0, 0) = G^T diag(-\hat{\boldsymbol{\lambda}}_0) X,$$

As $S_{\boldsymbol{\omega}}$ is a quadratic form, from standard statistical theory it has an asymptotic χ^2_q distribution, where q= $dim(\boldsymbol{\gamma})$, the dimension of $\boldsymbol{\gamma}$. In the case of a constant model for $\boldsymbol{\omega}$, this test reduces to that given by van den Broek (1995), more details see Appendix B. In our study, as we assume that $\boldsymbol{\omega} = \gamma_0$, q=1.

Chapter 3

Misspecification of Nuisance Parameters

3.1 Misspecification

As many authors pointed out, both the validity and the power of the test may be affected when the nuisance parameter is misspecified. In this report, we studied the impact of misspecifications of nuisance parameter, which can be described as below.

Consider a general statistical model represented by the log-likelihood function $L(\boldsymbol{\gamma}^*, \boldsymbol{\beta}^*)$, where $\boldsymbol{\gamma}^*$ and $\boldsymbol{\beta}^*$ are $q \times 1$ and $p \times 1$ vectors of parameters, respectively. Suppose $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, ...\beta_p^*)^T$, and $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, ...\tilde{\beta}_k)^T$, where $\boldsymbol{\beta}^*$ is a vector of true parameters and $\tilde{\boldsymbol{\beta}}$ is a $k \times 1$ vector of parameters other than $\boldsymbol{\beta}^*$. Then under the alternative, three types of misspecification of $\boldsymbol{\beta}$ are given as follows,

(1) $\boldsymbol{\beta}$ is a subset of the true parameters $\boldsymbol{\beta}^*$.

$$\boldsymbol{\beta} \subset (\beta_1^*, \beta_2^*, ...\beta_p^*)^T,$$

(2) $\boldsymbol{\beta}$ is contaminated. For example, $\boldsymbol{\beta} = (\beta_1^*, \beta_2^*, \tilde{\beta}_1, \tilde{\beta}_2)$, which means that $\boldsymbol{\beta}$ includes

the parameters that should not be included.

(3) $\boldsymbol{\beta}$ is totally misspecified.

$$\boldsymbol{\beta} \subseteq (\tilde{\beta}_1, \tilde{\beta}_2, ...\tilde{\beta}_k)^T,$$

In this report, we focus on these 3 types of misspecification and study the impact of misspecification of the Poisson mean on the power of score test for homogeneity in ZIP models.

3.2 Limiting distribution of $S_{\boldsymbol{\omega}}$ under misspecification

The score test statistic $S_{\boldsymbol{\omega}}$ is a quadratic form and from standard statistical theory it has an asymptotic ${\chi_q}^2$ distribution under the null hypothesis, see Jansakul and Hinde (2002). In order to investigate its null limiting distribution under misspecifications, we first conducted a simulation study using the score test proposed by van den Broek (1995).

The explanatory variables are: x_1, a continuous variable with truncated normal N(0,1) distributed values on (-1,1); x_2, a two level factor with two-fifths of the observations in the first group; x_3, a continuous variable with truncated normal N(1,1.5) distributed values on (0,2). We generated x_1 and x_3 from a multivariate normal distribution to make them orthogonal by setting the covariance between x_1 and x_3 equals zero. The true Poisson mean model is $\lambda^* = \exp(0.8 - 0.1x_1 + 0.3x_3)$ and the working models are specified in Table 3.1.

As shown in Table 3.1, when the Poisson mean is misspecified, the score test doesn't maintain the size as sample size increasing, i.e. n=800 and 1000, which indicates that the limiting distribution of the test statistics under the null no longer follows a χ^2 distribution.

Table 3.1: *The empirical sizes of S_ω at $\chi^2_{1;0.05}$ based on 1000 samples when the Poisson model is misspecified*

	Working Poisson		$\lambda^* = \exp(0.8 - 0.1x_1 + 0.3x_3)$			
	mean model	ω^*	n=50	n=200	n=800	n=1000
λ	$\beta_0 + \beta_1 x_1 + \beta_3 x_3$	$\omega^* = 0$	0.038	0.047	0.047	0.048
misspecified λ	β_0	$\omega^* = 0$	0.053	0.054	0.110	0.146
	$\beta_0 + \beta_1 x_1$	$\omega^* = 0$	0.053	0.059	0.104	0.117
	$\beta_0 + \beta_3 x_3$	$\omega^* = 0$	0.043	0.051	0.068	0.123
	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.051	0.049	0.118	0.124

[1] $x_1 \sim N(0,1)$ *and* $x_1 \in [-1,1]$, $x_2 \sim Bin(1, 0.6)$, $x_3 \sim N(1, 1.5)$ *and* $x_3 \in [0,2]$.

3.3 Parametric bootstrap

In practice, it is difficult to derive the true null limiting distribution of the score test statistic under the misspecification. However, a parametric bootstrap resampling method, which was first proposed by Efron (1979), can be used to find the true null limiting distribution. The bootstrap resampling method is often used to estimate distributions which are difficult to obtain analytically. It consists of 3 steps: (i) an estimation step, estimate the parameters of null model from the observed data; (ii) a Monte Carlo step, generate M pseudo-data sets from the fitted model and calculate the associate test statistics; finally, (iii) constructing distribution, construct the bootstrap distribution for a sufficient large value of M. Here we give the details of how we use this methodology to generate the large sample distribution of the score test statistic $S_{\boldsymbol{\omega}}$.

(1) Estimation step: compute the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$ under the null model for the given observed data $(y_i, x_i)_{i=1}^n$, where y_i are count outcome and x_i are covariates.

(2) Monte Carlo step: for each m, generate the Monte Carlo sample $(y_i^{(m)})_{i=1}^n$ from the null model where $\boldsymbol{\beta}$ fixed at $\hat{\boldsymbol{\beta}}$, assign each generated data point $y_i^{(m)}$ to x_i. Then for each

Monte Carlo sample $(y_i^{(m)}, x_i)_{i=1}^n$, calculate the score test statistic

$$S_{\boldsymbol{\omega}_n}^{(m)} = S_{\boldsymbol{\gamma}}{}^T(\hat{\boldsymbol{\beta}}_0^{(m)}, 0)\hat{C}^{-1}S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0^{(m)}, 0),$$

where

$$\hat{C} = J_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0^{(m)}, 0) - J_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0^{(m)}, 0)^T J_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0^{(m)}, 0)^{-1} J_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0^{(m)}, 0),$$

and $\hat{\boldsymbol{\beta}}_0^{(m)}$ is the estimate of $\hat{\boldsymbol{\beta}}$ under the null hypothesis using each generated data $(y_i^{(m)}, x_i)$, i=1,2...,n.

(3) Repeat step 2 for m=1,2...,M.

As M going large, an approximate p-value of score test can be calculated as

$$P_B = M^{-1}\sum_{m=1}^{M} I(S_{\boldsymbol{\omega}_n}^{(m)} \geq S_{obs}), \tag{3.1}$$

which is the proportion that the number of $S_{\boldsymbol{\omega}_n}^{(m)}$ greater than S_{obs}. We reject H_0 when P_B is smaller than the nominal value. In our simulation studies, we set M equal to 1000 and nominal value=0.05.

Chapter 4

Numeric Study

4.1 Simulation study

In the previous section 3.3, we mentioned that a bootstrap resampling method can be used to find the true null limiting distribution of the score test statistic when nuisance parameter is misspecified. In this section, by using this methodology, we investigated the effect of misspecified nuisance parameter on the power of the score test, specially the impact of misspecification of the Poisson mean. An intensive simulation study was carried out using R. In our simulations, we generated various samples of size n=25, 50, 100 and 200. For each data generating mechanisms and working models, we simulated 1000 sets of data from the true models. For each data set, we first calculated the observed S_{obs} values for some assumed working models by using the estimates from fitting the null model and then constructed its bootstrap distribution. The true models and various working models that we studied in this report are given in Table 4.1. In our simulations, we assume working ω is constant. The explanatory variables are: x_1, a continuous variable with truncated normal N(0,1) distributed values on (-1,1); x_2, a two level factor with two-fifths of the observations in the first group; x_3, a continuous variable with truncated normal N(1,1.5) distributed values on (0,2); x_4, a continuous variable with uniformly distributed values on (1,2). We generate

x_1 and x_3 from a multivariate normal distribution to make them orthogonal by setting the covariance between x_1 and x_3 equals zero.

Table 4.1: *Data generating mechanisms and working models*

True Models		Working Models for λ	
$\log(\lambda^*)$	ω^*	λ depends on	$\log(\lambda)$
0.7	0	constant	β_0
	0.1	1 covariate	$\beta_0 + \beta_1 x_1$
	$0.15 - 0.1x_1$		$\beta_0 + \beta_3 x_3$
$0.8 - 0.1x_3$	0	2 covariates	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
	0.1		$\beta_0 + \beta_1 x_1 + \beta_3 x_3$
	$0.15 - 0.1x_1$		$\beta_0 + \beta_2 x_2 + \beta_3 x_3$
$0.8 - 0.1x_1 + 0.3x_3$	0		$\beta_0 + \beta_2 x_2 + \beta_4 x_4$
	0.1	3 covariates	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
	$0.15 - 0.1x_1$	4 covariates	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

[1] $x_1 \sim N(0,1)$ *and* $x_1 \in [-1,1]$, $x_2 \sim Bin(1, 0.6)$,
$x_3 \sim N(1, 1.5)$ *and* $x_3 \in [0,2]$, $x_4 \sim U(1,2)$.

In order to check the size of $S_{\boldsymbol{\omega}}$, 1000 sets of data were generated from the null model $Y_i \sim Pois(\lambda_i)$, i=1,...n, where λ depends on the same covariates in the working models. For each data set, we first calculated the observed S_{obs} values by using the estimates from fitting the null model and constructed its bootstrap distribution. Then we calculated its p-value using equation (3.1). Finally, we calculated the proportion of times the p-value smaller than the critical value α, it can be written as

$$\frac{\sum_{m=1}^{1000} I(P_B^{(m)} < \alpha)}{1000}, \tag{4.1}$$

In our simulation, we set α=0.05.

Similarly, to investigate the impact of misspecification of the Poisson mean on the power of the test, we simulated 1000 sets of data from the true model $Y \sim ZIP(\lambda, \omega)$. For

each data set, we calculated the each observed S_{obs} value and its bootstrap distribution by using estimates from fitting the working models. Then we calculated p-value for each data set by using equation (3.1) and computed the power of the test using equation (4.1). For example, to investigate the power of the test under the true model ZIP(λ, ω), where $\lambda^* = \exp(0.8 - 0.1x_1 + 0.3x_3)$ and $\omega^* = 0.15 - 0.1x_1$, we first generated 1000 sets of data from this true model. Then for each data set, we calculated the each observed S_{obs} value and its bootstrap distribution by using estimates from fitting the assumed working models which are described in Table 4.1. Finally, we calculated the p-value by using equation (3.1) and computed the power of the test using equation (4.1).

4.2 Score test statistic under misspecification of Poisson mean

The results are presented in Tables 4.2, 4.3 and 4.4. From these tables we can see that,

(1) In Table 4.2, the sizes of the tests are all around 0.05. When the Poisson mean is misspecified (where the true Poisson mean is a constant), the power of the test decreases as more covariates are incorporated into the Poisson mean, where those covariates actually should not be included. Similarly, when ω^* depends on covariates, the power of the test also declines when more covariates are added into the Poisson mean, for example, when $\omega^* = 0.15 - 0.1x_1$ and n=50, the power decreases from 0.425 to 0.291 with the number of covariates in the Poisson mean increasing. However, the misspecification of the Poisson mean tends to have less impact on the power when sample size is large, for example, when $\omega^* = 0.15 - 0.1x_1$ and n=200, the power declines slightly from 0.949 to 0.920 as the number of covariates in the Poisson mean increases.

(2) In Table 4.3 and Table 4.4, we have three types of misspecification of the Poisson mean: 1) excluding the covariates that should be included, which refers to the first type that we mention in chapter 3; 2) including the covariates that should not be included, which

Table 4.2: *The empirical sizes and powers for S_ω at $\alpha = 0.05$ based on 1000 samples when $\lambda^* = \exp(0.7)$*

		$\lambda^* = \exp(0.7)$			
$\log\lambda$	ω^*	n=25	n=50	n=100	n=200
β_0	$\omega^* = 0$	0.045	0.053	0.054	0.049
	$\omega^* = 0.1$	0.171	0.252	0.451	0.767
	$\omega^* = 0.15 - 0.1x_1$	0.246	0.425	0.739	0.949
$\beta_0 + \beta_1 x_1$	$\omega^* = 0$	0.050	0.052	0.055	0.051
	$\omega^* = 0.1$	0.149	0.225	0.446	0.738
	$\omega^* = 0.15 - 0.1x_1$	0.205	0.406	0.700	0.931
$\beta_0 + \beta_3 x_3$	$\omega^* = 0$	0.054	0.043	0.040	0.047
	$\omega^* = 0.1$	0.134	0.243	0.426	0.726
	$\omega^* = 0.15 - 0.1x_1$	0.223	0.392	0.713	0.944
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.053	0.050	0.052	0.051
	$\omega^* = 0.1$	0.114	0.213	0.406	0.710
	$\omega^* = 0.15 - 0.1x_1$	0.188	0.379	0.650	0.938
$\beta_0 + \beta_1 x_1 + \beta_3 x_3$	$\omega^* = 0$	0.055	0.041	0.057	0.051
	$\omega^* = 0.1$	0.120	0.211	0.386	0.710
	$\omega^* = 0.15 - 0.1x_1$	0.178	0.358	0.643	0.940
$\beta_0 + \beta_2 x_2 + \beta_3 x_3$	$\omega^* = 0$	0.057	0.050	0.052	0.052
	$\omega^* = 0.1$	0.121	0.209	0.413	0.714
	$\omega^* = 0.15 - 0.1x_1$	0.202	0.383	0.676	0.940
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.056	0.048	0.054	0.049
$+\beta_3 x_3$	$\omega^* = 0.1$	0.112	0.191	0.377	0.706
	$\omega^* = 0.15 - 0.1x_1$	0.167	0.309	0.633	0.921
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.051	0.048	0.053	0.051
$+\beta_3 x_3 + \beta_4 x_4$	$\omega^* = 0.1$	0.095	0.173	0.362	0.675
	$\omega^* = 0.15 - 0.1x_1$	0.128	0.291	0.593	0.920

[1] $x_1 \sim N(0,1)$ *and* $x_1 \in [-1,1]$, $x_2 \sim Bin(1, 0.6)$,
$x_3 \sim N(1, 1.5)$ *and* $x_3 \in [0,2]$, $x_4 \sim U(1,2)$.

refers to the second type; 3) totally misspecified, which refers to the third type. Both of the tables showed that the sizes of the tests are stable and around the nominal level $\alpha = 0.05$. For the first type of misspecification, the test with a constant Poisson mean has a better power, even though the Poisson mean truly depend on covariates. For example, in Table 4.3 the true $\lambda^* = \exp(0.8 - 0.1x_3)$, $\omega^* = 0.15 - 0.1x_1$ and n=50, the test gains the power from 0.384 to 0.438 when the Poisson mean leaves out the covariate x_3, which should be included.

Table 4.3: *The empirical sizes and powers for S_ω at $\alpha = 0.05$ based on 1000 samples when $\lambda^* = \exp(0.8 - 0.1x_3)$*

		$\lambda^* = \exp(0.8 - 0.1x_3)$			
$\log\lambda$	ω^*	n=25	n=50	n=100	n=200
β_0	$\omega^* = 0$	0.040	0.050	0.045	0.050
	$\omega^* = 0.1$	0.170	0.260	0.484	0.756
	$\omega^* = 0.15 - 0.1x_1$	0.235	0.438	0.716	0.937
$\beta_0 + \beta_1 x_1$	$\omega^* = 0$	0.053	0.049	0.048	0.047
	$\omega^* = 0.1$	0.139	0.239	0.447	0.742
	$\omega^* = 0.15 - 0.1x_1$	0.177	0.355	0.651	0.923
$\beta_0 + \beta_3 x_3$	$\omega^* = 0$	0.055	0.054	0.050	0.049
	$\omega^* = 0.1$	0.135	0.235	0.459	0.724
	$\omega^* = 0.15 - 0.1x_1$	0.195	0.384	0.673	0.932
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.053	0.055	0.049	0.048
	$\omega^* = 0.1$	0.113	0.212	0.416	0.708
	$\omega^* = 0.15 - 0.1x_1$	0.167	0.302	0.623	0.914
$\beta_0 + \beta_1 x_1 + \beta_3 x_3$	$\omega^* = 0$	0.050	0.049	0.045	0.044
	$\omega^* = 0.1$	0.123	0.215	0.416	0.710
	$\omega^* = 0.15 - 0.1x_1$	0.169	0.331	0.623	0.904
$\beta_0 + \beta_2 x_2 + \beta_3 x_3$	$\omega^* = 0$	0.050	0.051	0.052	0.048
	$\omega^* = 0.1$	0.121	0.222	0.396	0.719
	$\omega^* = 0.15 - 0.1x_1$	0.166	0.332	0.633	0.916
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.051	0.051	0.050	0.053
$+\beta_3 x_3$	$\omega^* = 0.1$	0.109	0.180	0.379	0.701
	$\omega^* = 0.15 - 0.1x_1$	0.142	0.278	0.552	0.894
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.054	0.053	0.054	0.041
$+\beta_3 x_3 + \beta_4 x_4$	$\omega^* = 0.1$	0.097	0.168	0.348	0.681
	$\omega^* = 0.15 - 0.1x_1$	0.136	0.229	0.539	0.882

[1] $x_1 \sim N(0,1)$ *and* $x_1 \in [-1,1]$, $x_2 \sim Bin(1, 0.6)$,
$x_3 \sim N(1, 1.5)$ *and* $x_3 \in [0,2]$, $x_4 \sim U(1,2)$.

However, when the second type of misspecification occurs, the power of the test decreases with the number of covariates in the Poisson mean increasing. For example, in Table 4.3, when $\omega^* = 0.15 - 0.1x_1$ and n=50, the power decreases from 0.384 to 0.229 when the Poisson mean incorporates additional covariates that should not be included, such as x_1, x_2 and x_4. Under the third type of misspecification, the performance of the test only depends on the number of covariates in the Poisson mean model, even when the mean is misspecified. For

Table 4.4: *The empirical sizes and powers for S_ω at $\alpha = 0.05$ based on 1000 samples when* $\lambda^* = \exp(0.8 - 0.1x_1 + 0.3x_3)$

		$\lambda^* = \exp(0.8 - 0.1x_1 + 0.3x_3)$			
$\log \lambda$	ω^*	n=25	n=50	n=100	n=200
β_0	$\omega^* = 0$	0.057	0.049	0.048	0.047
	$\omega^* = 0.1$	0.474	0.703	0.915	0.994
	$\omega^* = 0.15 - 0.1x_1$	0.704	0.900	0.993	1.000
$\beta_0 + \beta_1 x_1$	$\omega^* = 0$	0.051	0.054	0.049	0.052
	$\omega^* = 0.1$	0.433	0.660	0.912	0.993
	$\omega^* = 0.15 - 0.1x_1$	0.614	0.851	0.991	1.000
$\beta_0 + \beta_3 x_3$	$\omega^* = 0$	0.055	0.055	0.047	0.049
	$\omega^* = 0.1$	0.423	0.645	0.878	0.989
	$\omega^* = 0.15 - 0.1x_1$	0.580	0.845	0.990	1.000
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.055	0.046	0.053	0.048
	$\omega^* = 0.1$	0.392	0.639	0.879	0.993
	$\omega^* = 0.15 - 0.1x_1$	0.526	0.820	0.989	1.000
$\beta_0 + \beta_1 x_1 + \beta_3 x_3$	$\omega^* = 0$	0.049	0.049	0.049	0.050
	$\omega^* = 0.1$	0.351	0.607	0.862	0.987
	$\omega^* = 0.15 - 0.1x_1$	0.538	0.801	0.971	1.000
$\beta_0 + \beta_2 x_2 + \beta_3 x_3$	$\omega^* = 0$	0.042	0.056	0.052	0.051
	$\omega^* = 0.1$	0.364	0.594	0.876	0.988
	$\omega^* = 0.15 - 0.1x_1$	0.538	0.825	0.987	1.000
$\beta_0 + \beta_2 x_2 + \beta_4 x_4$	$\omega^* = 0$	0.057	0.050	0.053	0.054
	$\omega^* = 0.1$	0.408	0.614	0.879	0.987
	$\omega^* = 0.15 - 0.1x_1$	0.557	0.835	0.990	1.000
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.043	0.054	0.049	0.051
$+\beta_3 x_3$	$\omega^* = 0.1$	0.336	0.573	0.862	0.983
	$\omega^* = 0.15 - 0.1x_1$	0.475	0.787	0.972	1.000
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	$\omega^* = 0$	0.048	0.051	0.047	0.053
$+\beta_3 x_3 + \beta_4 x_4$	$\omega^* = 0.1$	0.297	0.539	0.831	0.979
	$\omega^* = 0.15 - 0.1x_1$	0.414	0.768	0.966	1.000

[1] $x_1 \sim N(0,1)$ *and* $x_1 \in [-1,1]$, $x_2 \sim Bin(1, 0.6)$,
$x_3 \sim N(1, 1.5)$ *and* $x_3 \in [0,2]$, $x_4 \sim U(1,2)$.

example, in Table 4.4, the true Poisson mean is $\lambda^* = \exp(0.8 - 0.1x_1 + 0.3x_3)$ and when λ^* is totally misspecified as $\lambda = \exp(\beta_0 + \beta_2 x_2 + \beta_4 x_4)$, with $\omega^* = 0.15 - 0.1x_1$ and n=50, the power of the test is equal to 0.835, which is slightly higher than the power that is obtained under the well-specified model, which equals 0.801. We can also see this interesting result in Table 4.3 where the true $\lambda^* = \exp(0.8 - 0.1x_3)$. When λ^* is totally misspecified as $\lambda = \exp(\beta_0 + \beta_1 x_1)$, with $\omega^* = 0.15 - 0.1x_1$ and n=50, the power of the test is equal to 0.355 which is also slightly higher than the power that is obtained under the well-specified

model, which equals 0.384. This result gives us a strong evidence that the power of the test is only affected by the number of the covariates in the Poisson mean, regardless of the misspecification.

In sum, when the Poisson mean is misspecified, the power of the test decreases as the number of covariates in the Poisson mean increases. When sample size is large, the power of the test decreases slightly as the number of the covariates in the Poisson mean increases. However, no matter the Poisson mean is specified or not, the test with a constant Poisson mean is more powerful than other tests assumed the Poisson mean depends on covariates. This interesting finding gives us a suggestion that we can conduct the test by assuming a constant Poisson mean when evaluating homogeneity in ZIP models, in other words, we assume the Poisson mean model doesn't depend on any covariates. Besides, it is surprising that the power of the test is only affected by the number of the covariates in the Poisson mean.

4.3 Applications to Wuhan Inpatient Care Insurance data

We now illustrate the use of these findings with Wuhan Inpatient Care Insurance example.

4.3.1 Wuhan Inpatient Care Insurance data

The data in this study is obtained from Dongxihu District, Wuhan, the China Life Insurance Company. These data were collected by insurance salesmen and claims staff of this company between 1 July 2012 and 30 June 2013. The sample size is n=1562.

The outcome variable is the number of claims (NC) (ranges from 0 to 6). The distribution of NC for 1562 observations is given in Figure 4.1. Here, the 'zeros' constituted 90.01% of the observations.

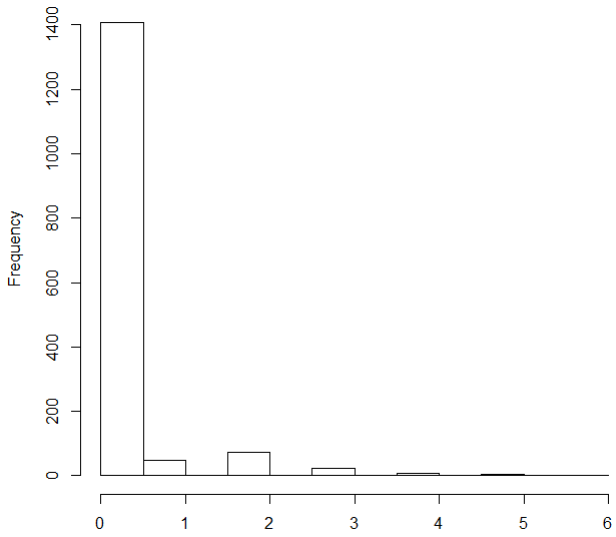

Figure 4.1: *Histogram of the number of claims*

Generally, NC usually contains excess zeros and is potentially affected by the factors of insured person's demographic characteristics (gender, age, marital status, education, monthly income, height, weight), occupation (labor or non-labor), living habits (smoking, drinking) and others (BMI, hereditary disease). Only hereditary disease is excluded since all the 1562 observations have no hereditary disease, thus eleven potential factors are used in our study. The details of these potential factors are:

`Gender` (=0 for male, and 1 for female), `Age` (ranges from 2 to 70 years old), `Education` (categorized as 0=none, junior high school and below, 1=senior high school including secondary and vocational school, 2=junior college, and 3=undergraduate, graduate and above), `Occupation` (dichotomized as 0=non-labor type, consisting of manager, doctor and nurse, teacher, civil, financial professionals, IT professionals, technician, business staff, administrative staff, self employed households, and others; and 1=labor type, including driver and conductor, construction site foreman, and worker), `Marital status` (dichotomized as 0=single, consisting of separated, divorced, widowed and never married person; and 1=currently married), `Monthly income` (in Chinese Yuan (RMB), categorized as 0=no income, 1=1-2,000 RMB, 2=2,001-5,000 RMB, 3=5,001-10,000 RMB, 4=10,001-20,000 RMB, 5=more than

20,000 RMB), `Height` (in centimeter (cm), ranges from 90 to 190), `Weight` (in kilogram (kg), ranges from 12 to 125), `BMI` (body mass index, calculated by BMI=Weight(kg)/Height(m)^2 and `Smoking` (=0 for nonsmokers and 1 for smokers), `Drinking` (=0 for nondrinkers and 1 for drinkers).

4.3.2 Testing result

First, we conducted a set of score tests to test whether the homogeneous Poisson regression is adequate or not for these data, see Table 4.5.

Table 4.5: *The score test statistics under different working models*

$\log\lambda$	S_{obs}	$p-value$
β_0	751.64	< 0.001
$\beta_0 + \beta_1 * \texttt{Gender}$	655.65	< 0.001
$\beta_0 + \beta_5 * \texttt{Marital status}$	593.34	< 0.001
$\beta_0 + \beta_2 * \texttt{Age}$	592.77	< 0.001
$\beta_0 + \beta_1 * \texttt{Gender} + \beta_5 * \texttt{Marital status}$	512.41	< 0.001
$\beta_0 + \beta_2 * \texttt{Age} + \beta_5 * \texttt{Marital status}$	410.85	< 0.001
$\beta_0 + \beta_1 * \texttt{Gender} + \beta_2 * \texttt{Age} + \beta_5 * \texttt{Marital status}$	360.87	< 0.001
$\beta_0 + \beta_1 * \texttt{Gender} + \beta_2 * \texttt{Age} + \beta_4 * \texttt{Occupation}$ $+\beta_5 * \texttt{Marital status} + \beta_7 * \texttt{Height} + \beta_9 * \texttt{BMI}$	359.57	< 0.001
Full model[1]	356.09	< 0.001
$\beta_0 + \beta_1 * \texttt{Gender} + \beta_2 * \texttt{Age} + \beta_5 * \texttt{Marital status} + \beta_9 * \texttt{BMI}$	356.08	< 0.001
$\beta_0 + \beta_1 * \texttt{Gender} + \beta_2 * \texttt{Age}$	345.93	< 0.001
$\beta_0 + \beta_1 * \texttt{Gender} + \beta_2 * \texttt{Age} + \beta_4 * \texttt{Occupation}$ $+\beta_5 * \texttt{Marital status} + \beta_9 * \texttt{BMI}$	342.67	< 0.001

[1] Full model = $\beta_0+\beta_1*\texttt{Gender}+\beta_2*\texttt{Age}+\beta_3*\texttt{Education}+\beta_4*\texttt{Occupation}+\beta_5*\texttt{Marital status}+\beta_6*\texttt{Monthly income}+\beta_7*\texttt{Height}+\beta_8*\texttt{Weight}+\beta_9*\texttt{BMI}+\beta_{10}*\texttt{Smoking}+\beta_{11}*\texttt{Drinking}$.

Table 4.5 suggests that the homogeneous Poisson regression is not adequate no matter

how the Poisson mean is specified, for these test statistics have significant p-values, which are less than 0.001.

4.3.3 Modeling results

Table 4.6: *Results of fitting a full ZIP model with ICI data*

Poisson component coefficients (Poisson with log link)			
	Estimate	Standard Error	p-value
Intercept	2.6771	14.1539	0.8500
`Gender`	0.6139	0.4329	0.1562
`Age`	-0.0225	0.0096	0.0193 *
`Education`	-0.0513	0.1148	0.6548
`Occupation`	0.0717	0.3170	0.8212
`Marital status`	0.3031	0.2599	0.2435
`Monthly income`	-0.2369	0.1201	0.0486 *
`Height`	-0.0104	0.0867	0.9045
`Weight`	0.0549	0.1592	0.7302
`BMI`	-0.1308	0.4340	0.7630
`Smoking`	-0.6567	0.3092	0.0337 *
`Drinking`	0.3837	0.2195	0.0805
Zero component coefficients (binomial with logit link)			
	Estimate	Standard Error	p-value
Intercept	-6.0471	11.6604	0.6040
`Gender`	-0.0719	0.4941	0.8843
`Age`	-0.1043	0.0170	< 0.001 **
`Education`	-0.0467	0.1456	0.7483
`Occupation`	0.0415	0.3111	0.8939
`Marital status`	-0.1791	0.2999	0.5503
`Monthly income`	-0.3450	0.1778	0.0523
`Height`	0.0650	0.0714	0.3626
`Weight`	-0.0433	0.1314	0.7416
`BMI`	0.2093	0.3642	0.5655
`Smoking`	-0.9155	0.4190	0.0289 *
`Drinking`	0.4182	0.3058	0.1714

[1] **:p-value< 0.01, *:p-value< 0.05.

The testing results in previous section 4.3.2 showed that the homogeneous Poisson model doesn't fit the ICI data adequately. In order to investigate which covariates affect the number of claims significantly, we use the ZIP model to analyze the Inpatient Care Insurance data

in our study. The ZIP model we used is introduced in chapter 2 with the link function $\log(\boldsymbol{\lambda}) = X\boldsymbol{\beta}$ and $\log\left(\frac{\boldsymbol{\omega}}{1-\boldsymbol{\omega}}\right) = G\boldsymbol{\gamma}$, where X and G are covariate matrices and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ are vectors of parameters. We first fit a full ZIP model using all covariates. The results are shown in Table 4.6, it is clear that many covariates are not significant. So we go on to fit several ZIP models using different combinations of the 11 covariates to find out which factors have the most impact on the number of claims, see Table 4.7.

Table 4.7: *Fits of the different ZIP models*

No.		model	AIC
1	$\log(\lambda)$	All covariates	1338.8
	ω	All covariates	
2	$\log(\lambda)$	Gender + Age + Education + Occupation + Marital status +Monthly income + Height + Weight + BMI + Smoking + Drinking	1328
	ω	Gender + Age + Marital status + Monthly income + BMI +Smoking + Drinking	
3	$\log(\lambda)$	Age + Marital status + Monthly income +Marital status*Monthly income + Smoking + Drinking	1326.4
	ω	Age + BMI + Smoking + Drinking	
4	$\log(\lambda)$	Age + Marital status + Monthly income + Smoking + Drinking	1326.27
	ω	Age + Monthly income + BMI + Smoking + Drinking	
5	$\log(\lambda)$	Gender + Age + Marital status + Monthly income + Smoking +Drinking	1325.4
	ω	Gender + Age + Monthly income + BMI + Smoking + Drinking	
6	$\log(\lambda)$	Age + Marital status + Monthly income + Age*Marital status +Marital status*Monthly income + Smoking + Drinking	1325.4
	ω	Age + BMI + Smoking + Drinking	

[1] All covariates = Gender + Age + Education + Occupation + Marital status + Monthly income + Height + Weight + BMI + Smoking + Drinking.

Here we consider the well-known AIC (Akaike information criterion) as a model selection criterion. In general, AIC is

$$AIC = 2k - 2\log L,$$

where k is the number of parameters in the model and L is the maximized value of the

likelihood function for the estimated model.

From Table 4.7, it is easy to see that AIC suggests the models 5 and 6 to be the most appropriate model with the smallest AIC value=1325.4. We choose model 6 as the final model because it is less complicated than the model 5. The estimated coefficients for model 6 are given in Table 4.8. Clearly, the Poisson mean λ depends significantly on `Age`, `Marital status`, `Monthly income`, `Smoking` and `Drinking` and the mixing weight ω depends significantly on `Age`, `BMI`, `Smoking` and `Drinking`. Note that the final model is selected based on AIC criterion, but we can not grantee the model 6 is the best model for the data.

Table 4.8: *Results of fitting the chosen ZIP model with ICI data*

Poisson component coefficients (Poisson with log link)			
	Estimate	Standard Error	p-value
Intercept	1.2453	0.5714	0.0293 *
`Age`	-0.0664	0.0268	0.0132 *
`Marital status`	0.4990	0.7546	0.5084
`Monthly income`	0.3386	0.2063	0.1007
`Age*Marital status`	0.0494	0.0282	0.0805
`Marital status*Monthly income`	-0.5485	0.2268	0.0156 *
`Smoking`	-0.8789	0.2893	0.0023 **
`Drinking`	0.3710	0.1618	0.0218 *
Zero component model coefficients (binomial with logit link)			
	Estimate	Standard Error	p-value
Intercept	3.0758	0.9086	0.0007 **
`Age`	-0.1065	0.0133	< 0.001 **
`BMI`	0.1076	0.0427	0.0118 *
`Smoking`	-1.0819	0.4095	0.0082 **
`Drinking`	0.7094	0.2498	0.0045 **

[1] **:p-value< 0.01, *:p-value< 0.05.

In Table 4.8, from the Poisson component we can see that for those married persons, as the monthly income increases, they tend to have less claims. Moreover, people who like to drink are seem to claim more times than those nondrinkers. In addition, both the age and smoking habit have negative effect on the NC, which means that the elder people and smokers tend to have less number of claims.

Chapter 5

Discussions

Through an intensive simulation study, we show that when the Poisson mean is misspecified, the limiting distribution of the score test statistic under the null is not a χ^2 distribution. This may be due to the violation of the regularity assumptions for score test. Moreover, the power of the test decreases as the number of covariates in the Poisson mean increases, regardless the mean is well specified or not. And the test for homogeneity has the better power when the working Poisson mean model doesn't depend on any covariates. This interesting finding suggests that we can conduct a test for homogeneity in ZIP models by assuming a constant Poisson mean, in other words, assuming the Poisson mean doesn't depend on any covariates.

In the real data analysis we have two main goals: 1) to evaluate that whether a standard Poisson can fit the Wuhan Inpatient Care Insurance data adequately; 2) to determine what covariates affect the number of claims significantly. For the first one, the test result shows that the standard Poisson regression doesn't fit the data well because of excess zeros. For the second one, we select a final model from several ZIP models with different sets of covariates by using the well-known AIC as a model selection criterion. The final model suggests that there are six factors affecting the number of claims significantly: age, marital status, monthly income, BMI, smoking and drinking habit.

There are some open questions that are subject to future research. For example, the find-

ings in this report are all obtained by simulation studies. The rigorous analytical evidences are still needed to support these findings. Another example is that we only investigate the impact of misspecification on the performance of score test for the homogeneity in ZIP models, ones can examine the same issue in other zero-inflated models, for example, the zero-inflated binomial (ZIB) model and the zero-inflated negative binomial (ZINB) model.

Bibliography

[1] Aerts, M., Claeskens, G., Hart, J.D. (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association*, 94(447), 869-879.

[2] Bera, A.K., Yoon, M.J. (1993). Specification testing with locally misspecified alternatives. *Econometric Theory*, 9, 649-658.

[3] Dietz, E., Böhning, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and Data Analysis*, 34, 441-459.

[4] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1-26.

[5] Godfrey, L.G. *Misspecification tests in econometrics: the Lagrange multiplier principle and other approaches*. Cambridge University Press, 1988.

[6] Jansakul, N., Hinde, J. P. (2002). Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis*, 40, 75-96.

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

[8] Liang, K.Y., Self, S.G. (1996). On the asymptotic behaviour of the pesudolikelihood ratio test statistic. *Journal of the Royal Statistical Society. Series B*, 58(4), 785-796.

[9] Molenberghs, G., Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician*, 61(1), 22-27.

[10] Todem, D., Hsu, W.W. (2012). On the efficiency of score tests for homogeneity in two-component parametric models for discrete data. *Biometrics*, 68(3), 975-982.

[11] van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, 51, 738-743.

Appendix A

Likelihood Ratio Test and Wald Test

Within the family of ZIP models, testing whether a Poisson model is adequate or not corresponds to testing the mixing weight ω

$$H_0 : \boldsymbol{\omega} = 0 \qquad vs \qquad H_1 : \boldsymbol{\omega} > 0, \tag{A.1}$$

where possible test statistics are the likelihood ratio test (LRT), the Wald test and the score test. However, the LRT and the Wald test statistics require the model under the alternative hypothesis to be estimated. For a general ZIP regression model, the LRT for zero-inflation is given by

$$LRT_{\boldsymbol{\omega}} = -2 \times [l(\boldsymbol{\lambda}_0) - l(\boldsymbol{\lambda}, \boldsymbol{\omega})],$$

where $l(\boldsymbol{\lambda}_0)$ and $l(\boldsymbol{\lambda}_0, \boldsymbol{\omega})$ are the maximized log-likelihoods under the Poisson regression and the ZIP regression models, respectively. The corresponding Wald test statistic is

$$W_{\boldsymbol{\omega}} = \boldsymbol{\omega}^T \Big[Cov(\boldsymbol{\omega}) \Big]^{-1} \boldsymbol{\omega},$$

which, in the case of single constant ω parameter, simplifies to

$$W_{\boldsymbol{\omega}} = \frac{\omega^2}{Var(\omega)},$$

Standard asymptotic theory would suggest that under H_0 both LRT_ω and W_ω are χ^2_1 distributed. However, for the ZIP model, the null hypothesis corresponds to $\boldsymbol{\omega}$ being on the boundary of the parameter space and the appropriate reference distribution is a mixture of χ^2 distributions, see Liang and Self (1987) and Feng and McCulloch (1992). For the simple constant ω model, the appropriate reference distribution is an equal mixture of a χ^2_0 (a constant at zero) and a χ^2_1 distribution, with p-value given by $\frac{1}{2}[Pr(\chi^2_1 \geqslant W_\omega)]$, etc.

A.1 Reference

Feng, Z., McCulloch, C.E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistic Probability Letter*, 13, 325-332.

Self, S.G., Liang, K.Y. (1987). Asymptotic properties of the maximum likelihood estimators and likelihood ratio test under nonstandard conditions. *Journal of American Statistic Association*, 82, 605-610.

Appendix B

Jansakul and Hinde's General Score Test Statistics

Based on the log-likelihood function given in (2.3) and the general model equations (2.5), the score vector is:

$$S(\boldsymbol{\beta};\boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{S_\beta}(\boldsymbol{\beta},\boldsymbol{\gamma}) \\ \boldsymbol{S_\gamma}(\boldsymbol{\beta},\boldsymbol{\gamma}) \end{bmatrix} = \begin{bmatrix} \frac{\partial l(\boldsymbol{\lambda},\boldsymbol{\omega})}{\partial \boldsymbol{\beta}} \\ \frac{\partial l(\boldsymbol{\lambda},\boldsymbol{\omega})}{\partial \boldsymbol{\gamma}} \end{bmatrix}$$

where,

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \lambda_i}\frac{\partial \lambda_i}{\partial \beta_j} = \sum_{i=1}^{n}[I_(y_i = 0)[\frac{-(1-\omega)e^{-\lambda_i}}{\omega_i + (1-\omega_i)e^{-\lambda_i}}]\lambda_i + I_(y_i > 0)(y_i - \lambda_i)]x_{ij}, j = 1, 2, ..., p$$

and

$$\frac{\partial l}{\partial \gamma_r} = \frac{\partial l}{\partial \omega_i}\frac{\partial \omega_i}{\partial \gamma_r} = \sum_{i=1}^{n}[I_(y_i = 0)\frac{(1-e^{-\lambda_i})}{\omega_i + (1-\omega_i)e^{-\lambda_i}} + I_(y_i > 0)(\frac{-1}{1-\omega_i})]g_{ir}, r = 1, 2, ..., q$$

The expected information matrix $J(\boldsymbol{\beta}, \boldsymbol{\gamma})$ can be partitioned as

$$J(\boldsymbol{\beta}; \boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{J}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) & \boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ \boldsymbol{J}_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) & \boldsymbol{J}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \end{bmatrix}$$

where the element $J_{\boldsymbol{\beta}}$, $J_{\boldsymbol{\beta}\boldsymbol{\gamma}} = J_{\boldsymbol{\gamma}\boldsymbol{\beta}}{}^T$ and $J_{\boldsymbol{\gamma}}$ are, respectively

$$-E\Big[\frac{\partial^2 l(\boldsymbol{\gamma}, \boldsymbol{\omega})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\Big], -E\Big[\frac{\partial^2 l(\boldsymbol{\gamma}, \boldsymbol{\omega})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}}\Big], and - E\Big[\frac{\partial^2 l(\boldsymbol{\gamma}, \boldsymbol{\omega})}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T}\Big],$$

with

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} \Bigg\{ I_{(y_i=0)} \left[\frac{-e^{-\lambda_i} \left[(1-\lambda_i)\omega_i + (1-\omega_i)e^{-\lambda_i} \right] (1-\omega_i)\lambda_i}{\left[\omega_i + (1-\omega_i)e^{-\lambda_i} \right]^2} \right] + I_{(y_i>0)}(-\lambda_i) \Bigg\} x_{ij} x_{ik}, j, k = 1, 2, \ldots, p,$$

$$\frac{\partial^2 l}{\partial \gamma_r \partial \gamma_s} = \sum_{i=1}^{n} \left\{ I_{(y_i=0)} \left[\frac{-(1-e^{-\lambda_i})^2}{\left[\omega_i + (1-\omega_i)e^{-\lambda_i} \right]^2} \right] + I_{(y_i>0)} \left[\frac{-1}{(1-\omega_i)^2} \right] \right\} g_{ir} g_{is},$$

$$r, s = 1, 2, \ldots, q$$

and

$$\frac{\partial^2 l}{\partial \gamma_r \partial \gamma_s} = \sum_{i=1}^{n} [I_(y_i = 0)[\frac{-1 - e^{-\lambda_i}{}^2}{\omega_i + (1-\omega_i)e^{-\lambda_i}{}^2}]] x_{ij} g_{ir},$$

Under the null hypothesis, the general score test is then

$$S_{\boldsymbol{\omega}} = S_{\boldsymbol{\gamma}}{}^T(\hat{\boldsymbol{\beta}}_0, 0) C^{-1} S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0), \tag{B.1}$$

where $\boldsymbol{\beta}_0$ is the maximum likelihood estimate under the Poisson model and

$$S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0) = G^T \Big[\frac{I_{y_i=0} - e^{-\hat{\boldsymbol{\lambda}}_0}}{e^{-\hat{\boldsymbol{\lambda}}_0}} \Big], \tag{B.2}$$

$$C = J_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0) - J_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0)^T J_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0, 0)^{-1} J_{\boldsymbol{\beta}\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0), \tag{B.3}$$

with

$$J_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0, 0) = X^T diag(\hat{\boldsymbol{\lambda}}_0) X, \tag{B.4}$$

$$J_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0) = G^T diag(\frac{1 - e^{-\hat{\boldsymbol{\lambda}}_0}}{e^{-\hat{\boldsymbol{\lambda}}_0}}) G, \tag{B.5}$$

and

$$J_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0, 0) = G^T diag(-\hat{\boldsymbol{\lambda}}_0) X, \tag{B.6}$$

In the case of a constant model for $\boldsymbol{\omega}$ this test reduces to that given by van den Broek (1995), i.e. if G is taken to be an $n \times 1$ matrix of 1's, then

$$S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0) = \sum_{i=1}^{n} \Big[\frac{I_{y_i=0} - e^{-\hat{\lambda}_{0i}}}{e^{-\hat{\lambda}_{0i}}} \Big],$$

$$J_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_0, 0) = \sum_{i=1}^{n} \Big[\frac{1 - e^{-\hat{\lambda}_{0i}}}{e^{-\hat{\lambda}_{0i}}} \Big],$$

and

$$J_{\boldsymbol{\gamma}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_0, 0) = -\hat{\boldsymbol{\lambda}}_o^T X,$$

The score test for a ZIP model with constant $\boldsymbol{\omega}$ is then

$$\frac{\Big[\sum_{i=0}^{n} \Big(\frac{I_{y_i=0} - \exp(-\hat{\lambda}_{0i})}{\exp(-\hat{\lambda}_{0i})} \Big) \Big]^2}{\Big[\sum_{i=0}^{n} \Big(\frac{1 - \exp(-\hat{\lambda}_{0i})}{\exp(-\hat{\lambda}_{0i})} \Big) \Big] - \hat{\boldsymbol{\lambda}}_0^T X \Big[X^T diag(\hat{\boldsymbol{\lambda}}_0) X \Big]^{-1} X^T \hat{\boldsymbol{\lambda}}_0},$$

which is equivalent to expression (3) in van den Broek (1995). Note that in this simple case,

the score statistics simply compare the observed zero frequency with the expected value under the Poisson model with appropriate weights.

Appendix C

R Code Example

```
library(truncnorm)
library(tmvtnorm)

ZIP=function(n0,p2,a,a0,b1,b2,b3,b4,b5,b6)
 {
 #initialize i,seq,x1,x2,x4,x5
 i=0
 seq=numeric(0)
mvn=rtmvt(1,c(0,1),diag(c(1,1)),lower=c(-1,1),upper=c(1,2))
x1=c(mvn[,1])
 x2=rbinom(1,n0,p2)
 x3=c(mvn[,2])
 x4=runif(1,1,2)
 x5=rbinom(1,n0,p2)
#p depend on covariate
p=a0
#u depend on covariate
u=exp(b3)
#generate 1 count
 cp=p+(1-p)*exp(-u)*(u^0)/(factorial(0))
 x=runif(1,0,1)
 while(cp<=x){
 py=(1-p)*exp(-u)*(u^(i+1))/(factorial(i+1))
```

```
 cp=cp+py
 i=i+1
 }
 seq=c(seq,i,p,u,x1,x2,x3,x5)
 return(seq)
 }
#generate ZIP of size n
 ZIPR=function(n,n0,p2,a,a0,b1,b2,b3,b4,b5,b6)
 {
 i=1
 seq1=numeric(0)
 while(i<=n){
 z=c(ZIP(n0,p2,a,a0,b1,b2,b3,b4,b5,b6))
 seq1=c(seq1,z)
 i=i+1
 mat=matrix(seq1,7)
 }
 return(mat)
 }
test=function(M,n,n0,p2,a,a0,b1,b2,b3,b4,b5,b6)
{
#initialize dataset, covariate
a=ZIPR(n,n0,p2,a,a0,b1,b2,b3,b4,b5,b6)
x=cbind(matrix(1,n,1))
g=cbind(matrix(1,n,1))
y=t(t(a[1,]))
x40=t(t(a[4,]))
x50=t(t(a[5,]))
x60=t(t(a[6,]))
seq2=matrix(,1,M)
#estimated beta0
m=glm(y~1, family="poisson")
b0=m$coef
#lamda
lamda=exp(x%*%b0)
#diagnal matrix of lamda
```

```
D=diag(c(lamda))
#score
s1=t(g)%*%(((y==0)*1-exp(-lamda))/exp(-lamda))
#information matrix
J11=t(x)%*%D%*%x
J22=t(g)%*%diag(c(((1-exp(-lamda))/exp(-lamda))))%*%g
J21=t(g)%*%diag(c(-lamda))%*%x
J12=t(J21)
C=J22-J21%*%solve(J11)%*%J12
#test statistics
sobs=t(s1)%*%solve(C)%*%s1
sobs=c(sobs)
for(jj in 1:M){
#bootstrap
y1=rpois(n,exp(x%*%b0))
m=glm(y1~1, family="poisson")
b=m$coef
#lamda
lamda=exp(x%*%b)
#diagnal matrix of lamda
D=diag(c(lamda))
#score
s1=t(g)%*%(((y1==0)*1-exp(-lamda))/exp(-lamda))
#information matrix
J11=t(x)%*%D%*%x
J22=t(g)%*%diag(c(((1-exp(-lamda))/exp(-lamda))))%*%g
J21=t(g)%*%diag(c(-lamda))%*%x
J12=t(J21)
C=J22-J21%*%solve(J11)%*%J12
#test statistics
s=t(s1)%*%solve(C)%*%s1
s=c(s)
seq2[1,jj]=s
}
s95=quantile(seq2,c(0.95))
S0=(sobs>s95)*1
```

```
return(S0)
}
#calculate N test statistics and proportion
testR=function(N,M,n,n0,p2,a,a0,b1,b2,b3,b4,b5,b6,alpha,q)
{
i=1
seq3=numeric(0)
while(i<=N){
z1=c(test(M,n,n0,p2,a,a0,b1,b2,b3,b4,b5,b6))
seq3=c(seq3,z1)
i=i+1
}
P=mean(seq3)
return(P)
}
#b0,0.05
testR(1000,1000,50,1,0.6,1,0,0,0,0.7,0,0,0,0.05,1)
testR(1000,1000,50,1,0.6,1,0.1,0,0,0.7,0,0,0,0.05,1)
```

Appendix D

中国人寿保险股份有限公司
武汉市分公司东西湖支公司
中国 湖北 武汉市东西湖区东吴大道 1142 号
China Life Insurance Company Limited
1142 Dongwu Av., Dongxihu District, Wuhan 430040, Hubei, P.R.China
Tel: +86-027-83215521

中国人寿
CHINA LIFE

Certificate

Nov. 28, 2013

This is to certify that all data are from insured persons' insurance and insured claims data of the China Life Insurance Company Limited. The data is real and valid, only as scientific research purposes, not for commercial purposes.

人寿保险股份有限公司
武汉市分公司
(非订立新合同章)

Wuhan Branch of China Life Insurance Company Limited