Chin Med Sci J
December 2017

Vol. 32, No. 4
P. 218-225

CHINESE
MEDICAL SCIENCES
JOURNAL

ORIGINAL ARTICLE

Study of Zero-Inflated Regression Models in a Large-Scale Population Survey of Sub-Health Status and Its Influencing Factors△

Tao Xu[1], Guangjin Zhu[2], Shaomei Han[1]*

[1]Department of Epidemiology and Statistics, [2]Department of physiopathology, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College, Beijing 100005, China

Key words: zero-inflated; negative binomial regression; sub-health; population survey

Objective Sub-health status has progressively gained more attention from both medical professionals and the publics. Treating the number of sub-health symptoms as count data rather than dichotomous data helps to completely and accurately analyze findings in sub-healthy population. This study aims to compare the goodness of fit for count outcome models to identify the optimum model for sub-health study.

Methods The sample of the study derived from a large-scale population survey on physiological and psychological constants from 2007 to 2011 in 4 provinces and 2 autonomous regions in China. We constructed four count outcome models using SAS: Poisson model, negative binomial (NB) model, zero-inflated Poisson (ZIP) model and zero-inflated negative binomial (ZINB) model. The number of sub-health symptoms was used as the main outcome measure. The alpha dispersion parameter and *O* test were used to identify over-dispersed data, and Vuong test was used to evaluate the excessive zero count. The goodness of fit of regression models were determined by predictive probability curves and statistics of likelihood ratio test.

Results Of all 78307 respondents, 38.53% reported no sub-health symptoms. The mean number of sub-health symptoms was 2.98, and the standard deviation was 3.72. The statistic *O* in over-dispersion test was 720.995 (P<0.001); the estimated alpha was 0.618 (95% *CI*: 0.600-0.636) comparing ZINB model and ZIP model; Vuong test statistic Z was 45.487. These results indicated over-dispersion of the data and excessive zero counts in this sub-health study. ZINB model had the largest log likelihood (-167519), the smallest Akaike's Information Criterion coefficient (335112) and the smallest Bayesian information criterion coefficient (335455), indicating its best goodness of fit. The predictive probabilities for most counts in ZINB model fitted the observed counts best. The logit section of ZINB model analysis showed that age, sex, occupation, smoking, alcohol drinking, ethnicity and obesity were determinants for presence of sub-health symptoms; the binomial negative

Received: March 30, 2017.

*Corresponding author Tel: 86-10-69156408, E-mail: hansm1@vip. sina. com

△Fund supported by the Basic Performance Key Project, the Ministry of Science and Technology of the People's Republic of China (No. 2006FY110300).

section of ZINB model analysis showed that sex, occupation, smoking, alcohol drinking, ethnicity, marital status and obesity had significant effect on the severity of sub-health.

Conclusions All tests for goodness of fit and the predictive probability curve produced the same finding that ZINB model was the optimum model for exploring the influencing factors of sub-health symptoms.

Chin Med Sci J 2017; 32(4):218-225. DOI:10.24920/J1001-9294.2017.054

RECENTLY sub-health becomes an important public health issue and has attracted more and more attention from both medical professionals and the public. However, people usually think that sub-health is a borderline state between health and disease, which might be caused by strong social stress. Sub-health is a suboptimal health status, and an intermediate health state between health and disease, which is characterized by a decline in vitality, physiological function and capacity for adaptation, and it refers to medically undiagnosed or functional somatic syndromes.[1,2] It has been reported that 60%-70% of Chinese people suffering from sub-health.[3] It is very important to properly assess sub-health and explore the its influencing factors in order to prevent diseases and promote the health status of the general population.

Nowadays, sub-health status is usually assessed with a variety of rating scales. Delphi self-rating sub-health scale is very applicable to sub-health assessment for community population. It assesses 18 symptomatic items. If a subject has been suffering from one or more symptoms for more than one month in the past year, he or she will be considered as of sub-health status.[4] Previous studies showed that Delphi self-rating sub-health scale had good reliability and repeatability.[4-6]

In rating scales, subjects are usually categorized into two or several groups based on whether they had positive symptoms. The prevalence rate and logistic regression are usually used to study the incidence intensity and the relationship between sub-health and its risk factors.[6-10] However, these traditional methods may lose information, because every subject has different number of sub-health symptoms so that categorization will result in inability to assess the severity of sub-health status. Actually, the number of sub-health symptoms is a kind of count data, in which observations take only non-negative integer values {0, 1, 2, 3 ...}. During statistical treatment, if they are considered as continuous outcomes or transferred to dichotomous data, the numbers of sub-health symptoms are often extremely concentrated and don't follow the normal distribution. Consequently, arithmetic means and standard deviations are not applicable, and linear regression is not appropriate because of skewed distribution and over-dispersion. Although linear regression and logistic regression models are often used to treat count data, the results are likely to be inefficient, inconsistent or biased.[11,12]

In view of these limitations, Poisson regression or negative binomial (NB) regression is commonly used to model count outcomes based on their specific distributions. However, a lot of subjects have no sub-health symptom and there is a large proportion of zero count, which has adverse effect on the goodness of fit of Poisson regression or NB models. Neglect of excess zeroes will bias the estimation of parameters.[13] Zero-inflated (ZI) regression models consider the raw dataset as a mixture of an all-zero subset and another subset following Poisson or NB distribution.[13-20] ZI models were firstly introduced by Lambert to explain excess zero counts.[21] Cheung YB mentioned that ZI regression models can be interpreted as reckoning a two-step disease regression.[13] At the beginning subjects are not at the risk, so they have zero counts. The influence of covariates may move them into the at-risk population and the outcomes follow a Poisson or NB regression distribution. A covariate may or may not have the same direction of impact in the two steps.

This study was designed with a large-scale population survey to compare the goodness of fit of four count outcome models: Poisson model, NB model, Zero-inflated Poisson (ZIP) model and Zero-inflated negative binomial (ZINB) model, aiming at identifying the optimum model for the study of sub-health status.

MATERIALS AND METHODS

Sample and participants

A large-scale population survey about physiological and psychological constants was conducted in Chinese people during 2007-2011. It was supported by the basic performance key project of the Ministry of Science and Technology of the People's Republic of China. This survey was conducted in four provinces: Sichuan, Heilongjiang, Hunan, Yunnan; and two autonomous regions: Inner Mongolia, and Ningxia Hui. Two-stage cluster sampling method was used to select eligible subjects in every province. Firstly, two or three cities were sampled based