Too many zeros in your longitudinal data? Use zero-inflated mixed models

Wei-Wen Hsu*
Department of Statistics, Kansas State University
and
Emily L. Mailey
Department of Kinesiology, Kansas State University

August 15, 2018

Abstract

The simplest design of longitudinal study for a continuous outcome is that the response variable is only measured twice throughout the whole study. Typically, data are collected at the beginning and the end of the study. The classical statistical methods such as the paired t-test and repeated measures ANOVA models are often used to analyze such data in order to investigate whether the outcome variable changes over time or a difference exists between control and intervention groups over time. However, these classical methods could provide unreliable statistical inferences if ignoring that too many zeros are observed in the outcome variable. As a solution, zero-inflated mixed models is recommended for analyzing this type of data. In this article, we use the data collected from a web-based intervention program as an example to illustrate how to use zero-inflated mixed models for longitudinal data with excess zeros.

Keywords: Zero inflation, Continues outcome, linear mixed models, control intervention study.

*Email: wwhsu@ksu.edu

1 Introduction

For a longitudinal control-intervention study, it is very common to use the design in which the continuous outcome variable Y (usually a nonnegative variable, i.e. ≥ 0) is only measured twice across the whole study (i.e. measured at the beginning and the end of study). This design is often used to address the research question: "Is there a difference in change over time for the outcome variable between control and intervention groups?" In other words, the evaluation of efficacy of intervention is the main goal. This question in general can be answered with the independent t-test or repeated measures ANOVA models.

The independent t-test simply can examine whether there is a difference between two groups in the change of outcome over time. Although the important assumption for t-test is normality, it is generally not problematic with reasonably large samples. Repeated measures ANOVA models are also popular and widely used to address the same question in different areas of application. This type of model requires more assumptions. In addition to the assumptions of normality and independence between different subjects, an extra assumption called sphericity is given. This assumption is also known as the compound symmetric assumption which implies that the correlations among repeated measurements are equal and the variances at each of the repeated measurements are constant (Fitzmaurice et al., 2012; Twisk, 2013).

It is important to ensure all required assumptions are satisfied before using the above methods. However, there is a little attention in practical literatures to address the excess zeros observed in the longitudinal data while performing the data analysis. Especially for longitudinal continuous outcome variable, there is a few. But there are several studies established for longitudinal count data with excess zeros (see, for example, Wang et al., 2002; Yau et al., 2003; Min and Agresti, 2005; Lee et al., 2006).

In this article, we primarily focus on the situation where too many zeros are observed in the continuous outcome variable in a longitudinal study. As known, those extra zeros would result in the violation of normality and further fail the repeated measures ANOVA models. For such scenario, we suggest the use of zero-inflated mixed models to accommodate these extra zeros in order to answer the same research question mentioned earlier. As an example, the real data from a web-based intervention program study are used to illustrate the use of zero-inflated mixed models.

2 Zero-inflated mixed model for continuous outcome

We assume there are n independent subjects indexed by i ($i = 1, \cdots, n$), and within each subject i there are m_i repeated measurements indexed by j ($j = 1, \cdots, m_i$). The zero-Inflated mixed model for a continuous outcome is a two-component mixture model which combines a degenerate distribution at zero and a parametric non-degenerate distribution dominated by a set of parameters $\boldsymbol{\zeta}$. Specifically, the zero-inflated mixed model is defined as,

$$f(y_{ij}) = \begin{cases} \omega_{ij} & \text{if } y_{ij} = 0 \\ (1 - \omega_{ij})\, g(y_{ij}; \boldsymbol{\zeta}) & \text{if } y_{ij} \neq 0 \end{cases}$$

where $f(.)$ denotes the probability density function and Y_{ij} is the continuous outcome variable with observed value y_{ij} and ω_{ij} is the unknown mixture probability. The subject-specific effects can be captured by including random terms into ω_{ij} and/or the mean μ_{ij} of the parametric distribution $g(y_{ij}; \boldsymbol{\zeta})$. For example, it is very common to assume that

$$\log\left(\frac{\omega_{ij}}{1 - \omega_{ij}}\right) = Z_{ij}\gamma + u_i$$

and

$$\log(\mu_{ij}) = X_{ij}\beta + v_i$$

where Z_{ij} and X_{ij} are covariate matrices and γ, β are the associated coefficient vectors; u_i and v_i are subject-specific random effects. In general, u_i and v_i are assumed to be independent and normally distributed with mean zero and variance σ_u^2, σ_v^2, respectively. For simplicity, even though not realistic, the mixture can be assumed to be a constant cross all subjects, i.e. $\omega_{ij} = \omega$. In practice, this mixed model can be easily fitted using SAS with the procedure PROC NLMIXED.

3 Example: A web-based intervention to promote physical and mental health among military spouses

As an example, we use a subset of data from a web-based intervention study which was mainly designed to compare an interactive, theory-based web-delivered intervention to a more generic,

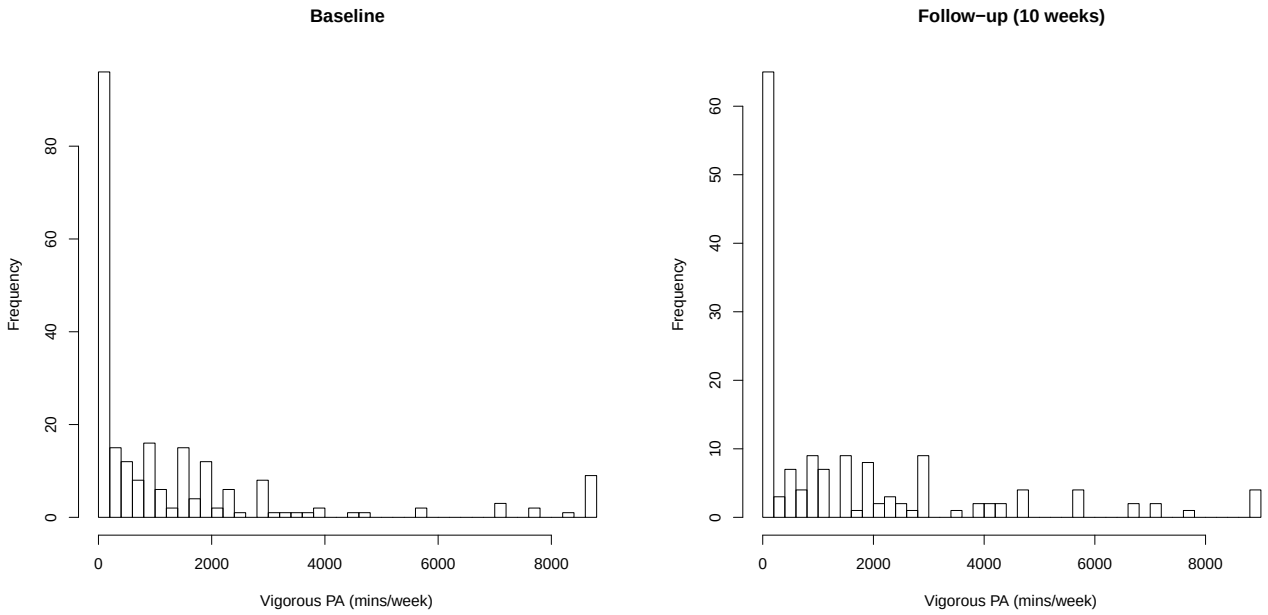

Figure 1: The histograms of the outcome variable at baseline and follow-up

educational web-delivered intervention (control). Participants were all military spouses. It was expected to see this interactive web-based intervention program can improve their mental and physical health outcomes significantly, compared to the control. In their study, the outcomes were only measured at the beginning and the end of intervention (i.e. two repeated measurements only).

Among many health outcomes, we select one outcome variable, Vigorous Physical Activity (PA), to illustrate the use of zero-inflated mixed model. In Figure 1, it is clear to see the existence of many zeros in the outcome variable at baseline and the follow-up (the proportions of zeros are 40.6% and 28.1%, respectively). Thus, the violation of normality is expected when using repeated measures ANOVA models. Unfortunately, there is no good data transformation for such data in order to satisfy the normality assumption, that is simply because of too many zeros.

As the first step of the data analysis, we transform the non-zero observations by using a log transformation and keep zeros as they are. Using such transformation is because we want to achieve the normality approximately for these non-zero observations. We then can assume that the transformed observations are generated from normal distribution with mean μ_{ij} and variance σ^2; in other words, we assume these non-zero observations are generated from a log-normal distribution. This step is basically to set up the parametric non-degenerate distribution in a zero-inflated model. Second, we assume a constant parameter to represent the probability of having the extra zero and

this probability is actually referred as the mixture probability in a zero-inflated model (i.e. $\omega_{ij} = \omega$ for all i, j). We further assume the overall mean of these transformed observations depends on certain covariates (or predictors): Age, Number of Children, Group (control or intervention) and Wave (baseline or follow-up). That is, $\mu_{ij} = \beta_0 + \beta_1 * Age + \beta_2 * (Number\ of\ Children) + \beta_3 * Group + \beta_4 * Wave + \beta_5 * Wave * Group$. We are particularly interested in the interaction term of Wave*Group, which can be evaluated to determine whether the intervention program is effective over time. It is worth to mention that we do not include a random effect term into the mean. Instead, we use the idea of Working Independence Model coupled with the use of sandwich estimator to do the parameter estimation. We use sandwich estimator to estimate the variance-covariance matrix in order to avoid any misspecification of covariance structure (i.e. the underlying correlations among repeated measurements). Misspecification of variance-covariance matrix will lead to unreliable inferences. The theoretical derivation and the idea of Working Independence Model can be found in Freedman (2006) and Hsu et al. (2014). In SAS, we could use the 'EMPIRICAL' option to compute the sandwich estimator. However, this option does require the specification of random effects terms in the model, which we do not specify previously. As a solution, using options 'QPOINTS=1' and 'NOAD' additionally can obtain the Working Independence Model likelihood and then provide the estimates of parameters and statistical inferences based on the sandwich estimator. The SAS example is given in Appendix.

4 The data analysis result

The results of the data analysis are given in Table 4.1. For Vigorous PA outcome, the mixture probability is significant (estimate = 0.416, p-value<0.001), which suggests the existence of zero inflation in the data and the use of zero-inflated models is appropriate. Unfortunately, there is no significance for the coefficient of interaction term (Wave*Group). There is no evidence to conclude the efficacy of intervention over time.

5 Discussion

The classical methods for longitudinal data are expected to fail when too many zeros are present in the outcome variable. That is simply because the normality assumption can not be satisfied.

Table 4.1: Estimated coefficients of zero-inflated mixed model for the Vigorous PA data

	Estimate	S.E.	p-value
Intercept	6.675	0.402	<0.001
Age(years)	0.017	0.014	0.239
Number of Children	0.005	0.042	0.908
Group(0=Control 1=intervention)	0.072	0.172	0.677
Wave(0=baseline 1=follow-up)	0.495	0.158	0.002
Group*Wave	−0.370	0.220	0.094
Probability of extra zero (ω)	0.416	0.029	<0.001
σ^2	0.845	0.069	<0.001

Also, there is often no proper data transformation can be used for zero inflation in order to allow classical methods such as repeated measures ANOVA models to analyze the data. Therefore, we recommend the zero-inflated mixed model coupled with the use of sandwich estimator for such situation, especially this model can accommodate the extra zeros and provide reliable statistical inferences. In practice, this model can be easily fitted using SAS procedure PROC NLMIXED with minimal programming efforts. It is worth to note that we assume a constant mixture probability in this study, which is often not realistic but just simple. Practical analysts can use a link function, for example a logistic link, to relate this mixture probability to covariates and subject-specific random effects. This approach can be also done in SAS easily (but not shown).

Appendix

The SAS code for the zero-inflated mixed model is given below. Here $y = \log(\text{Vigorous PA})$ is the outcome variable and the covariates for the mean are 'Age', 'Number of Children', 'Wave', 'Group' and the interaction 'Wave*Group'.

```
data long;
set IBNA_long_out;
if waves=2 then wave1=1;else wave1=0;
if Vigorous_pa=0 then y=0;
if Vigorous_pa~=0 then y=log(Vigorous_pa);
ngroup=group-1;
run;

proc nlmixed data=long empirical noad qpoints=1 maxiter=500;
parameters b0-b5=0.1 sigma2=0.1 wi=0.5;
bounds 0<=wi<=1;
eta2= b0 + b1*age + b2*ngroup + b3*Num_children +b4*wave1 + b5*wave1*ngroup + r;

/* Normal mean  */
mu=(eta2);

/* Build the ZINormal log likelihood */
if y=0 then
ll = log( wi );
else ll = log(1-wi) -0.5*log(6.28) - 0.5*log(sigma2) - 0.5*(y-mu)**2/sigma2;
model y ~ general(ll);
random r ~ normal(0,1) subject=id;
run;
```

References

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2012). *Applied longitudinal analysis*. John Wiley & Sons.

Freedman, D. A. (2006). On the so-called huber sandwich estimator and robust standard errors. *The American Statistician 60*(4), 299–302.

Hsu, W.-W., D. Todem, K. Kim, and W. Sohn (2014). A wald test for zero inflation and deflation for correlated count data from dental caries research. *Statistical Modelling 14*(6), 471–488.

Lee, A. H., K. Wang, J. A. Scott, K. K. Yau, and G. J. McLachlan (2006). Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros. *Statistical methods in medical research 15*(1), 47–61.

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

Twisk, J. W. (2013). *Applied longitudinal data analysis for epidemiology: a practical guide*. Cambridge University Press.

Wang, K., K. K. Yau, and A. H. Lee (2002). A zero-inflated poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. *Computer Methods and Programs in Biomedicine 68*(3), 195–203.

Yau, K. K., K. Wang, and A. H. Lee (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal 45*(4), 437–452.