Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

A time-varying effect model for examining group differences in trajectories of zero-inflated count outcomes with applications in substance abuse research

Songshan Yang[a], James A. Cranford[b], Jennifer M. Jester[b], Runze Li[c], Robert A. Zucker[b], Anne Buu[d]*

This study proposes a time-varying effect model for examining group differences in trajectories of zero-inflated count outcomes. The motivating example demonstrates that this zero-inflated Poisson model allows investigators to study group differences in different aspects of substance use (e.g. the probability of abstinence and the quantity of alcohol use) simultaneously. The simulation study shows that the accuracy of estimation of trajectory functions improves as the sample size increases; the accuracy under equal group sizes is only higher when the sample size is small (100). In terms of the performance of the hypothesis testing, the type I error rates are close to their corresponding significance levels under all settings. Furthermore, the power increases as the alternative hypothesis deviates more from the null hypothesis and the rate of this increasing trend is higher when the sample size is larger. Moreover, the hypothesis test for the group difference in the zero component tends to be less powerful than the test for the group difference in the Poisson component. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: Time-varying effect; zero-inflated Poisson model; substance abuse; longitudinal data; B-spline

1. Introduction

National data showed significant gender and racial/ethnic differences in substance use from early adolescence to young adulthood [1]. Characterizing group differences (such as gender or racial/ethnic differences) in developmental trajectories

[a] *Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.*
[b] *Department of Psychiatry & Addiction Research Center, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
[c] *Department of Statistics and The Methodology Center, Pennsylvania State University, University Park, PA 16802, U.S.A.*
[d] *Department of Health Behavior and Biological Sciences, University of Michigan, Ann Arbor, MI 48109, U.S.A.*
* *Correspondence to: Department of Health Behavior and Biological Sciences, University of Michigan, 400 North Ingalls Street, Ann Arbor, MI 48109, U.S.A. E-mail: buu@umich.edu*

Contract/grant sponsor: National Institutes of Health

of substance use has the potential to provide crucial information about the special timing and risky patterns of each group. Such information may be used to design more targeted prevention and intervention strategies.

Some important methodological issues arise as we pursue this line of research. First, the group difference may change in terms of both magnitude and direction across time. For example, a national longitudinal study found that Hispanic youth had higher rates of substance use in early adolescence, whereas Caucasian youth had higher levels of substance use from mid-adolescence through the early 30s; such ethnic differences largely disappeared after age 30 [1]. The commonly adopted growth curve model pre-specifies a simple shape for developmental changes and tests group differences implicitly through the interactions between the group indicator and linear/quadratic terms. Such simple shapes can hardly characterize the complex developmental trajectory of substance use for each group, especially when there are many time points spanning a long developmental period. Second, many outcomes of interest in the substance abuse field are count data with excess zeros such as the number of drinks consumed or the number of substance use symptoms reported [2]. Our previous work demonstrated that when excess zeros exist in the data, the conventional Poisson regression model tends to produce high mean squared error [3]. Furthermore, when the zero-inflated Poisson (ZIP) model is adopted to handle excess zeros, its performance tends to be better in the Poisson component than in the zero component in the setting of variable selection [2]. Given that group differences in developmental trajectories are examined in both components in the ZIP model, whether its relative performance in the two components varies with different situations is an open research question.

A recent study by our group [4] proposed a time-varying effect model (TVEM) that explicitly characterizes gender differences in developmental trajectories of substance use by modeling gender as a time-varying effect. Such trajectories are estimated through non-parametric regression functions that do not assume fixed shapes like conventional growth curves. We also demonstrated in the same paper that the TVEM model is very useful for (1) characterizing developmental changes across the life span based on multi-wave longitudinal studies; and (2) delineating patterns of health risk behaviors based on short-term studies collecting intensive longitudinal data such as daily process data or ecological momentary assessment data. Furthermore, the model can handle a variety of longitudinal outcomes under the framework of generalized linear models [5]. The aim of this paper is to extend our previous work to characterize group-specific trajectories of zero-inflated count outcomes and conduct hypothesis testing for group differences. We demonstrate the programming for carrying out the proposed ZIP model in SAS [6]. We also conduct simulations to evaluate the performance of the ZIP model in both the zero and Poisson components based on the data features of the Michigan Longitudinal Study (MLS), which is an ongoing multi-wave prospective study of youth at high risk for alcoholism. The alcohol use measure collected from this sample from childhood to adulthood provides a typical example of longitudinal zero-inflated count data.

This paper is organized as follows. In Section 2, we specify the TVEM for zero-inflated count data and propose the procedures to practically implement the estimation. In Section 3, we present a motivational example using the MLS data. In Section 4, we present results of simulation studies assessing the performance of the proposed model in terms of the accuracy of estimation of trajectories, type I error rate, and statistical power under different conditions. Discussion and concluding remarks are presented in Section 5. A SAS program example is given in the Appendix.

2. The Statistical Model

2.1. The zero-inflated Poisson (ZIP) model

The zero-inflated Poisson (ZIP) model was originally proposed to model the number of defects on an item in a manufacturing process that is assumed to move randomly back and forth between a perfect state (i.e. zero) and an imperfect state (i.e. Poisson) [7]. The model has been applied in many fields including the substance abuse field [3]. In a longitudinal setting, let Y_{ij} be the j-th observed outcome from the i-th subject at time t_{ij} $(i = 1, \ldots, n; j = 1, \ldots, J_i)$ and k be the group (e.g. gender) that Subject i belongs to $(k = 1, 2)$. The ZIP model is a finite mixture model:

$$Y_{ij} = \begin{cases} 0 & \text{with probability } \pi_{ij} \\ \text{Poisson } \lambda_{ij} & \text{with probability } 1-\pi_{ij} \end{cases}$$

Thus, the probability distribution is written as:

$$P(Y_{ij}=0) = \pi_{ij} + (1-\pi_{ij})e^{-\lambda_{ij}}$$

$$P(Y_{ij}=y_{ij}) = (1-\pi_{ij})e^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}/(y_{ij}!), y_{ij} = 1,2,3.....$$

The parameters π_{ij} and λ_{ij} can be modeled by

$$\text{logit}(\pi_{ij}) = \mu_1(t_{ij}) + \beta_1(t_{ij})I_{\{k=1\}} + a_i \tag{1}$$

$$\log(\lambda_{ij}) = \mu_2(t_{ij}) + \beta_2(t_{ij})I_{\{k=1\}} + b_i \tag{2}$$

In the zero component, $\mu_1(t_{ij})$ is the trajectory of the $k=2$ group; $\beta_1(t_{ij})$ delineates the time-varying difference between the two groups; and a_i is a normally distributed random effect with the variance σ_1^2. In the Poisson component, $\mu_2(t_{ij})$ is the trajectory of the $k=2$ group; $\beta_2(t_{ij})$ delineates the time-varying difference between the two groups; b_i is a normally distributed random effect with the variance σ_2^2; and the covariance between a_i and b_i is σ_{12}.

Suppose Y_{ij}, $j=1,\ldots,J_i$ are conditionally independent given the random effects a_i and b_i, the likelihood function of the i-th subject is

$$L_i = \int \prod_{j=1}^{J_i} P(Y_{ij}=0)^{I(Y_{ij}=0)} P(Y_{ij}=y_{ij})^{1-I(Y_{ij}=0)} dF(a_i, b_i, \sigma_1^2, \sigma_2^2, \sigma_{12})$$

where $F(\cdot)$ is the joint distribution function. Thus, the likelihood function for all the subjects is

$$L = \prod_{i=1}^{n} L_i.$$

2.2. *Parameter estimation and hypothesis testing*

The nonparametric functions $\mu_1(t)$, $\mu_2(t)$, $\beta_1(t)$, and $\beta_2(t)$ can be represented as linear combinations of basis expansions. We choose to use a spline basis to represent them as piecewise cubic functions. This requires the range of t to be divided into segments by multiple knots. The number of knots controls the amount of smoothing and can be chosen by goodness-of-fit statistics such as BIC. On each of several intervals defined by knots, the spline function is continuous and has continuous first and second derivatives at the knots. This allows any smooth shape to be approximated well if enough knots are used. Specifically, we use a B-spline basis [8], which can be automatically generated by most commonly used statistical software such as SAS and R for a given set of knots. Each knot's basis function is orthogonal to the other basis functions except for its closest knot neighbors. In order to avoid the issue of overfitting when too many knots are used, we adopt the approach of Shiyko and colleagues [9] that used a small number of equally spaced knots and treated the selection of the number of basis functions as a model selection problem.

After defining the basis functions using the B-spline formula, our model can be modified as follows:

$$\text{logit}(\pi_{ij}) = \sum_{s=1}^{S} \alpha_{1s}\phi_s(t_{ij}) + \sum_{s=1}^{S} \alpha_{2s}\phi_s(t_{ij})I_{\{k=1\}} + a_i \tag{3}$$

$$\log(\lambda_{ij}) = \sum_{s=1}^{S} \alpha_{3s}\phi_s(t_{ij}) + \sum_{s=1}^{S} \alpha_{4s}\phi_s(t_{ij})I_{\{k=1\}} + b_i \quad (4)$$

where $\phi_1(t_{ij}), \ldots, \phi_S(t_{ij})$ are known functions of time defined using the recursive B-spline formulas; $\alpha_{11}, \ldots, \alpha_{1S}$, $\alpha_{21}, \ldots, \alpha_{2S}$, $\alpha_{31}, \ldots, \alpha_{3S}$ and $\alpha_{41}, \ldots, \alpha_{4S}$ are the corresponding regression coefficients. Therefore, the zip model can be written as a generalized linear mixed model [10] of which the parameters can be estimated by maximizing the log-likelihood function. A SAS program example is provided in the Appendix to demonstrate how to use the PROC NLMIXED to carry out the computation.

In practice, researchers are interested in graphing group specific trajectories. Based on the fixed effects in Equations (3, 4), the trajectories for the $k = 2$ group are $\mu_1(t_{ij}) = \sum_{s=1}^{S} \alpha_{1s}\phi_s(t_{ij})$ in the zero component and $\mu_2(t_{ij}) = \sum_{s=1}^{S} \alpha_{3s}\phi_s(t_{ij})$ in the Poisson component. For the $k = 1$ group, the trajectories are $\mu_1(t_{ij}) + \beta_1(t_{ij}) = \sum_{s=1}^{S} \alpha_{1s}\phi_s(t_{ij}) + \sum_{s=1}^{S} \alpha_{2s}\phi_s(t_{ij})$ in the zero component and $\mu_2(t_{ij}) + \beta_2(t_{ij}) = \sum_{s=1}^{S} \alpha_{3s}\phi_s(t_{ij}) + \sum_{s=1}^{S} \alpha_{4s}\phi_s(t_{ij})$ in the Poisson component. Furthermore, the delta method can be used to estimate the variances of the estimated functions of these two groups in both the zero and Poisson components at any time t_{ij}. In this way, we can obtain the pointwise confidence intervals which can be plotted along with the estimated trajectories.

In addition to estimating the group specific trajectories, researchers are interested in testing whether there exist any group differences in the zero component or the Poisson component of the ZIP model. We formulate the hypothesis testing problems for the zero and Poisson components, respectively, as follows:

Test 1. $H_0 : \beta_1(t) = 0$ v.s. H_1: $\beta_1(t) \neq 0$

Test 2. H_0: $\beta_2(t) = 0$ v.s. H_1: $\beta_2(t) \neq 0$

In Test 1, under H_0, the two groups have the same trajectories in the zero component. Following the method described above, we can estimate $\mu_1(t)$ under H_0, as well as $\mu_1(t)$ and $\beta_1(t)$ under H_1. We can further evaluate the log-likelihood functions under H_0 and H_1 denoted by $\ell(H_0)$ and $\ell(H_1)$, respectively. The generalized likelihood ratio test(GLRT) for the hypothesis can thus be defined by $T = 2\{\ell(H_1) - \ell(H_0)\}$. Because $\beta_1(t)$ is nonparametric, the degree of freedom of T is unknown and we can only derive the empirical distribution of T by simulation. Following Cai and colleagues [11], we can conduct bootstrap sampling to estimate the p-value for the GLRT. The detail of this simulation procedure is described in Section 4.2. A similar procedure can be applied to conduct Test 2 that examines whether the two groups have the same trajectories in the Poisson component.

When the hypothesis testing result is significant, it means that the two groups do not have the same trajectories. This GLRT is like an overall F test in ANOVA. A significant result leads to further comparison of the group-specific trajectories with confidence intervals, which can provide more specific information about where the difference comes from.

3. A Motivating Example: The Michigan Longitudinal Study (MLS)

The MLS is an ongoing prospective study of people at high risk for substance abuse and disorder [12]. It is the developmentally earliest study currently extant and is also one of the longest running projects in the substance abuse field. We chose to use data from the MLS to demonstrate the application of the proposed method, because the study is highly influential and the features of the data are typical in the field. Our simulation (described in the next section) is also built upon the ZIP model fitted on this data set and thus may be highly applicable to the field. Furthermore, analyzing these data using the ZIP model allows us to to examine gender differences in developmental trajectories of the probability of abstinence as well as the quantity of alcohol use. The results may inform future prevention and intervention work.

The MLS recruited participant families using fathers' drunk driving conviction records and door-to-door community canvassing in a four-county area in mid-Michigan. All participants received extensive in-home assessments of their

substance use and related risk factors and consequences at baseline, and thereafter at 3-year intervals. The children of participant families were followed from early childhood to adulthood. During the critical developmental period of alcohol use onset and peak use (early adolescence to young adulthood), these children were assessed annually in order to measure drinking onset and patterns more accurately. In this study, we use longitudinal data (ages 14 to 24) from a sample of 696 children (70% males) for analysis. The maximum number of time points available is 15, although some participants may skip certain time points.

The goals of our empirical analysis are to (1) characterize gender-specific alcohol use behavior developmentally from early adolescence to young adulthood, and (2) test gender differences in developmental trajectories. In our analysis, the outcome at each time point is the self estimated number of drinks per drinking day in past month. The zero values are more than what would be expected from a classical Poisson regression model (see Figure 1). Thus, statistical models designed for handling zero-inflated data such as the ZIP model are needed. More importantly, the ZIP model makes it possible to model the probability of abstinence (the zero component) and the quantity of alcohol use (the Poisson component) simultaneously. Gender is treated as a time-varying effect through $\beta_1(t)$ and $\beta_2(t)$ in Models (1, 2). Using AIC and BIC, we choose 1 knot to approximate the trajectories. Figure 2 presents the inverse logit and exponential transformations of the fitted gender-specific trajectories in the zero and Poisson components, which can be easily interpreted as the probability of abstinence and quantity of alcohol use, respectively. Panels (a) and (c) show the developmental trajectories with the asymptotic pointwise confidence intervals (CI) for females and males in the zero component, respectively. Although the probability of abstinence decreases from ages 14 to 20 for both gender groups, the rate of change is faster for male. Panel (e) summaries such gender differences. Furthermore, the result of the generalized likelihood ratio test indicates that there are significant gender differences with $T = 21$ and the corresponding p-value estimated to be 0.001 by bootstrap sampling. Panels (b) and (d) in Figure 2 delineate the developmental trajectories with CI for females and males in the Poisson component, respectively; Panel (f) summaries the corresponding time-varying gender differences. Overall, males tend to drink more than females. The quantity of alcohol use increases during adolescence for both gender groups, but females reach the peak and start to cut down earlier (before age 20) than males. Moreover, the result of the generalized likelihood ratio test indicates that there are significant gender differences with $T = 104$ and the corresponding p-value estimated to be close to 0 by bootstrap sampling.

4. Simulation

4.1. The Accuracy of Estimation of Trajectory Functions

In this section, we evaluate the performance of the proposed method in terms of its accuracy for estimating trajectory functions under different situations. The response is a count variable which follows the ZIP model introduced in Section 2.1. We generate the response Y using the estimated functions of $\mu_1(t)$, $\mu_2(t)$, $\beta_1(t)$, $\beta_2(t)$, a_i and b_i from the fitted model of the MLS data described in Section 3. We manipulate two factors: (i) the sample size: $n = 100, 200$ and 400; and (ii) the gender ratio among all the participants: male/female=1 and male/female=7/3 (the latter simulates the ratio in the real data example). All the simulations are based on 1000 replicates.

The criterion for evaluating the accuracy of estimation of trajectory functions is the mean integrated squared error (MISE):

$$\text{MISE} = \frac{1}{1000}\sum_{r=1}^{1000}\sum_{i=1}^{n}\sum_{j=1}^{J_i}((\mu_m(t_{rij}) - \hat{\mu}_{rm}(t_{rij}))^2 + (\beta_m(t_{rij}) - \hat{\beta}_{rm}(t_{rij}))^2)/(h_m^2),$$

where $m = 1$ refers to the zero component; $m = 2$ refers to the Poisson component; and h_m is the corresponding range of the trajectory functions. We calculate the MISE for the zero component as well as the one for the Poisson component under each combination of the sample size and the gender ratio.

Table 1 shows the MISE and its empirical standard error (SE). When the sample size increases and the gender ratio is fixed, the accuracy of the proposed method improves as demonstrated by the decreasing MISE and SE. This trend is observed in the zero component as well as in the Poisson component. The effect of the gender ratio is, on the other hand, smaller and inconsistent. The increased accuracy under the balanced gender ratio $(1:1)$ only appears for the smallest sample size. When the sample size is larger ($n > 100$), such superiority disappears.

4.2. The Type I Error Rate and Power of the Hypothesis Testing

We also evaluate the performance of the hypothesis testing on gender differences in both the zero and Poisson components concerning the type I error rate and power. Our simulation considers two hypothesis tests: (i) $H_0 : \beta_1(t) = 0$ versus $H_1 : \beta_1(t) = \delta\hat{\beta}_1(t)$; and (ii) $H_0 : \beta_2(t) = 0$ versus $H_1 : \beta_2(t) = \delta\hat{\beta}_2(t)$, where $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ are the trajectory functions in the ZIP model fitted on the MLS data; and the value of δ is manipulated to reflect different levels of deviation from H_0. In this part of our simulation study, we only manipulate the sample size while keeping the gender ratio at 7:3 (i.e. the ratio in the MLS data). Because the proposed model is non-parametric, we can only derive the empirical distribution of the test statistic $T = 2\{\ell(H_1) - \ell(H_0)\}$ by simulation. Under each sample size, we generate $2,000$ data sets under H_0 and then conduct the hypothesis testing which results in $2,000$ values of T. Thus, the critical values of $T_{0.01}, T_{0.05}, T_{0.10}$, and $T_{0.25}$ are the 99th, 95th, 90th, and 75th percentiles from this empirical distribution of T, respectively. After obtaining the critical values under each sample size, we examine the effect of δ on the power of the test by taking a grid of δ over $(0, 1.5)$. Under each sample size, we generate 500 data sets for each value of δ and conduct the hypothesis testing on the data sets (using the four critical values) to examine the type I error rate and the power corresponding to the four values of the significance level α.

Table 2 shows the type I error rates (i.e. the values of power when $\delta = 0$), which are close to their corresponding significance levels under all settings. Since we use 500 data sets to compute the type I error rates, the 95% percent confidence intervals of the type I error rates are 0.01 ± 0.009, 0.05 ± 0.019, 0.10 ± 0.026 and 0.25 ± 0.038. Therefore, the simulation result in Table 2 demonstrates that our method controls the type I error rates pretty well. Figure 3 depicts the power as a function of δ and α under different sample sizes for the two hypothesis tests. As demonstrated by the figure, the power increases as δ increases and the rate of this increasing trend is higher when the sample size is larger. For example, in Test 1 (see the left panel), for the sample size of 100, all the power curves reach 1 when δ is larger than 1.5, whereas for the sample size of 400, all the power curves achieve 1 when δ is less than 0.8. Furthermore, comparing the left panel and the right panel in Figure 3, we observe that Test 1 (the group difference in the zero component) tends to be less powerful than Test 2 (the group difference in the Poisson component) when the sample size, δ, and α are fixed.

5. Discussion

This study proposes a time-varying effect model for examining group differences in trajectories of zero-inflated count outcomes. We extend our prior work beyond the generalized linear model cases, so that we can model group differences in different aspects of substance use (e.g. the probability of abstinence and the quantity of alcohol use) simultaneously. The design of our simulation study is unique because it represents the special features of a well-known longitudinal study on alcoholism risk so that the results can be generalizable to the substance abuse field. We also fill in the knowledge gap by comparing the power of the test in the zero component and the one in the Poisson component. Furthermore, the proposed model can be applied to not only multi-wave longitudinal studies like the MLS, but also short-term studies that involve intensive data collection such as daily process data [4, 13] and ecological momentary assessment (EMA) data [14, 15, 16, 17].

The simulation study shows that the accuracy of estimation of trajectory functions improves as the sample size increases. The effect of the gender ratio is, on the other hand, smaller and inconsistent. The increased accuracy under the balanced

gender ratio only appears for the smallest sample size. When the sample size is larger ($n > 100$), such superiority disappears. This result has an important practical implication because in many fields of health science, particularly the alcohol and substance abuse field, participants are largely male as reflected in the gender ratio in the MLS data. Thus, the proposed method is highly applicable in the field. In terms of the performance of the hypothesis testing, the type I error rates are close to their corresponding significance levels under all settings. Furthermore, the power increases as the alternative hypothesis deviates more from the null hypothesis and the rate of this increasing trend is higher when the sample size is larger. Moreover, the hypothesis test for the group difference in the zero component tends to be less powerful than the test for the group difference in the Poisson component. This result is consistent with our previous work on variable selection showing that the performance of ZIP model tends to be better in the Poisson component than in the zero component [2].

The real data analysis demonstrates the major strength of the proposed model that allows us to examine gender differences in terms of the probability of abstinence and the quantity of alcohol use simultaneously. Although the probability of abstinence decreases across time for both gender groups, the rate of change is faster for male. Furthermore, males tend to drink more than females overall. The quantity of alcohol use increases during adolescence for both gender groups, but females reach the peak and start to cut down earlier than males. This implies that females tend to "mature out" (due to psychological maturity, or family/job responsibilities) earlier than males.

In this work, we used B-splines to approximate the nonparametric time-varying coefficients in Models (1) and (2), and developed an estimation and hypothesis testing procedure for the time-varying coefficients. The resulting confidence intervals are pointwise confidence intervals without considering the bias due to the spline approximation to coefficient functions. Additionally, the p-value for our hypothesis testing procedure is derived by using bootstrap sampling. In practice, the B-splines method typically involves model selection for specifying knots so the p-value should be adjusted for the stochastic error inherited in the model selection procedure. Thus, one should be cautious in interpretation of significance and the corresponding p-value in practice. These have been issues for most existing estimation and hypothesis testing procedures for nonparametric models, although nonparametric models are very useful for exploratory analysis.

In longitudinal studies like the MLS, participants tend to have different patterns of skipping or rescheduling their assessments across waves. Thus, we designed our simulation to reflect this common phenomenon by sampling the individual-specific set of t (i.e. the age at each assessment) from the MLS data. Because of this design, we were not able to manipulate the number of time points in the simulation study. Nevertheless, we would like to provide a practical advice about future applications of the proposed model to longitudinal studies that may have sparse assessments. Although the TVEM has been applied mostly to analysis of intensive longitudinal data, it does not necessarily require many repeated measures from each subject. As demonstrated in Figure 1 of Fan & Li [18], the TVEM also applies to the setting in which the observed time points within an individual subject are very sparse, and yet the total time points aggregated across all subjects are dense in the time interval of interest because every subject is repeatedly measured at different time (e.g. ages). This implies that a *varying* assessment schedule is preferred for future applications of the proposed model.

Although the methodology proposed in this study was motivated by our research interest in gender differences, it can be applied to a variety of contexts that involve the comparison between two trajectories or change patterns. For example, the model can be used to compare the substance use patterns in a treatment group and a control group, so that the time-varying treatment effect can be well characterized. Future work may be needed to extend the methodology to handle settings with more than two groups such as comparing the effects of multiple treatment arms. Furthermore, this study focuses on the setting involving a single substance use outcome. Future studies may extend the model to handle multiple substance use behaviors that tend to co-occur such as alcohol use, nicotine use, and marijuana use [19, 20, 21].

It would be an interesting research topic to develop model diagnostic tools for examining whether the proposed model is appropriate for a particular data set, and whether one should consider other zero-inflated models such as the zero-inflated negative binomial model. Residual analysis has been a useful tool for model diagnostics in the context of linear regression analysis. Nevertheless, due to the presence of random effects a_i and b_i in Models (1) and (2), we found that it is very

challenging in obtaining the prediction of the random effects in order to calculate the residuals. Thus, the development of model diagnostic tools would be a good topic for future research.

Appendix. SAS program example for carrying out the proposed method

PROC TRANSREG is employed to construct piecewise cubic functions in Equations (3, 4) using the degree three (`DEGREE=3`) B-spline (`BSPLINE`) with 1 knot (`NKNOTS=1`), based on preliminary analysis and model selection. This results in $3+1+1$ terms, `age_0-age_4`, which are stored in the data set `basis` (`PREDICTED` refers to these new variables).

```
PROC TRANSREG DATA=mls;
     MODEL IDENTITY(drinkday)=BSPLINE(age / DEGREE=3 NKNOTS=1);
     OUTPUT OUT=basis PREDICTED;
     RUN;
```

The following data step merges together the resulting transformations of `age` with the original data set `mls`, which contains the ID number (`target`), the outcome (`drinkday`), and the indicator variable for gender (`male`).

```
DATA final;
     MERGE mls basis;
     KEEP target drinkday male age_0-age_4;
     RUN;
```

We use PROC NLMIXED to fit the ZIP model described in Section 2.1. The variables created in the program are defined as follows:

`linkp`: logit(π_{ij}) in Equation 3
`random_binary`: the random effect a_i in Equation 3
`VarBinary`: the variance of a_i, i.e., σ_1^2
`p0`: the probability of abstinence
`mu`: λ_{ij} in Equation 4
`random_poisson`: the random effect b_i in Equation 4
`VarPoisson`: the variance of b_i, i.e., σ_2^2
`ll`: the log-likelihood function

```
PROC NLMIXED DATA=final COV; /* COV: covariance matrix of a1-d5 */
    /* Set up initial values for parameters */
    PARAMETER  a1=0 a2=0 a3=0 a4=0 a5=0
               b1=0 b2=0 b3=0 b4=0 b5=0
               c1=0 c2=0 c3=0 c4=0 c5=0
               d1=0 d2=0 d3=0 d4=0 d5=0
               VarBinary=1 VarPoisson=1 ;
    /* Define the ZIP model */
    linkp=random_binary+a1*age_0+a2*age_1+a3*age_2+a4*age_3+a5*age_4+
          c1*male*age_0+c2*male*age_1+c3*male*age_2+c4*male*age_3+c5*male*age_4;
    p0=exp(linkp)/(1+exp(linkp));
    mu=exp(random_poisson+b1*age_0+b2*age_1+b3*age_2+b4*age_3+b5*age_4+
```



```
    d1*male*age_0+d2*male*age_1+d3*male*age_2+d4*male*age_3+d5*male*age_4);
/* Compute the log-likelihood function */
IF drinkday=0 THEN
    ll=log(p0+(1-p0)*exp(-mu));
ELSE ll=log((1-p0))-mu+drinkday*log(mu)-lgamma(drinkday+1);
MODEL drinkday~GENERAL(ll);
/* ParameterEstimates output estimates of a1-d5, VarBinary & VarPoisson */
/* FitStatistics output AIC, BIC, and log-likelihood */
/* CovMatParmEst output the covariance matrix of a1-d5 */
ODS OUTPUT ParameterEstimates=MyPar FitStatistics=MyFit
          CovMatParmEst=MyCov;
RANDOM random_binary random_poisson ~
      NORMAL([0,0],[VarBinary,0,VarPoisson]) SUBJECT=target;
RUN;
```

Acknowledgements

The work of Yang and Li was supported by the National Institutes of Health (NIH) [P50 DA010075, P50 DA036107 & R01 DA039854] and the National Science Foundation (NSF) [DMS 1512422]; The work of Cranford and Buu was supported by the NIH [R01 DA035183]; and the work of Jester and Zucker was supported by the NIH [R01 AA07065]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and NSF.

References

1. Chen P, Jacobson KC. Developmental trajectories of substance use from early adolescence to young adulthood: gender and racial/ethnic differences. *Journal of Adolescent Health* 2012; **50**:154-163. DOI: 10.1016/j.jadohealth.2011.05.013
2. Buu A, Johnson NJ, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine* 2011; **30**:2326-2340. DOI: 10.1002/sim.4268
3. Buu A, Li R, Tan X, Zucker RA. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine* 2012; **31**:4074-4086. DOI: 10.1002/sim.5510
4. Yang S, Cranford JA, Li R, Zucker RA, Buu A. Time-varying effect model for studying gender differences in health behavior. *Statistical Methods in Medical Research* in press.
5. McCullagh P, Nelder JA. *Generalized linear models.* Chapman & Hall/CRC: Boca Raton, FL, 2005.
6. SAS Institute Inc. *SAS/STAT 9.2 user's guide.* SAS Institute Inc: Cary, NC, 2008.
7. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**:1-13. DOI: 10.2307/1269547
8. de Boor C. *A Practical Guide to Splines.* Springer-Verlag: New York, NY, 1978.
9. Shiyko MP, Lanza ST, Tan X, Li R, Shiffman S. Using the time-varying effect model (TVEM) to examine dynamic associations between negative affect and self-confidence on smoking urges: Differences between successful quitters and relapsers. *Prevention Science* 2012; **13**:288-299. DOI: 10.1007/s11121-011-0264-z
10. Molenberghs G, Verbeke G. *Models for discrete longitudinal data.* Springer: New York, NY, 2005.
11. Cai Z, Fan J, Li R. Efficient estimation and inferences for varying-coefficient models. *Journal of American Statistical Association* 2000; **95**:888-902. DOI: 10.2307/2669472
12. Zucker RA, Ellis DA, Fitzgerald HE, Bingham CR, Sanford K. Other evidence for at least two alcoholisms II: life course variation in antisociality and heterogeneity of alcoholic outcome. *Development and Psychopathology* 1996; **8**:831-848.
13. Yang H, Cranford JA, Li R, Buu A. Two-stage model for time-varying effects of discrete longitudinal covariates with applications in analysis of daily process data. *Statistics in Medicine* 2015; **34**:571-581. DOI: 10.1002/sim.6368

14. Liu X, Li R, Lanza ST, Vasilenko S, Piper M. Understanding the role of cessation fatigue in the smoking cessation process. *Drug and Alcohol Dependence* 2013; **133**:548-555. DOI: 10.1016/j.drugalcdep.2013.07.025
15. Selya AS, Dierker LC, Rose JS, Hedeker D, Tan X, Li R, Mermelstein RJ. Time-varying effects of smoking quantity and nicotine dependence on adolescent smoking regularity. *Drug and Alcohol Dependence* 2013; **128**:230-237. DOI: 10.1016/j.drugalcdep.2012.08.026
16. Shiyko M, Naab P, Shiffman S, Li R. (2014). Modeling complexity of EMA data: time-varying lagged effects of negative affect on smoking urges for subgroups of nicotine addiction. *Nicotine & Tobacco Research* 2014; **16S2**:S144-S150. DOI: 10.1093/ntr/ntt109
17. Vasilenko S, Piper M, Lanza ST, Liu X, Yang J, Li R. Time-varying processes involved in smoking lapse in a randomized trial of smoking cessation therapies. *Nicotine & Tobacco Research* 2014; **16S2**:S135-S143. DOI: 10.1093/ntr/ntt185
18. Fan J, Li R. An overview on nonparametric and semiparametric techniques for longitudinal data. In *Frontiers of Statistics*, Fan J, Koul H (eds). Imperial College Press: London, 2006; 277-304.
19. Buu A, Dabrowska A, Mygrants M, Puttler LI, Jester JM, Zucker RA. Gender differences in the developmental risk for onset of alcohol, nicotine, and marijuana use and the effects of nicotine and marijuana use on alcohol outcomes. *Journal of Studies on Alcohol and Drugs* 2014; **75**:850-858.
20. Buu A, Dabrowska A, Heinze JE, Hsieh H, Zimmerman MA. Gender differences in the developmental trajectories of multiple substance use and the effect of nicotine and marijuana use on heavy drinking in a high risk sample. *Addictive Behaviors* 2015; **50**:6-12. DOI: 10.1016/j.addbeh.2015.06.015
21. Lin H, Jester JM, Buu A. The relationships of cigarette and alcohol use with the initiation, re-initiation, and persistence of cannabis use. *Journal of Studies on Alcohol and Drugs* 2016; **77**:113-120.

Table 1. MISE under different sample sizes and gender ratios.

		$n = 100$		$n = 200$		$n = 400$	
	Male:Female	7:3	1:1	7:3	1:1	7:3	1:1
Zero component	MISE	0.863	0.663	0.424	0.372	0.140	0.183
	SE	0.603	0.461	0.284	0.121	0.105	0.133
Poisson component	MISE	0.205	0.196	0.174	0.182	0.123	0.134
	SE	0.112	0.107	0.083	0.093	0.048	0.059

Table 2. Type I error rates under different sample sizes.

Test 1: the zero component			
	n=100	n=200	n=400
$\alpha = 0.01$	0.010	0.012	0.010
$\alpha = 0.05$	0.054	0.046	0.044
$\alpha = 0.10$	0.106	0.096	0.08
$\alpha = 0.25$	0.276	0.250	0.226
Test 2: the Poisson component			
	n=100	n=200	n=400
$\alpha = 0.01$	0.010	0.012	0.018
$\alpha = 0.05$	0.052	0.060	0.042
$\alpha = 0.10$	0.108	0.118	0.08
$\alpha = 0.25$	0.244	0.282	0.238

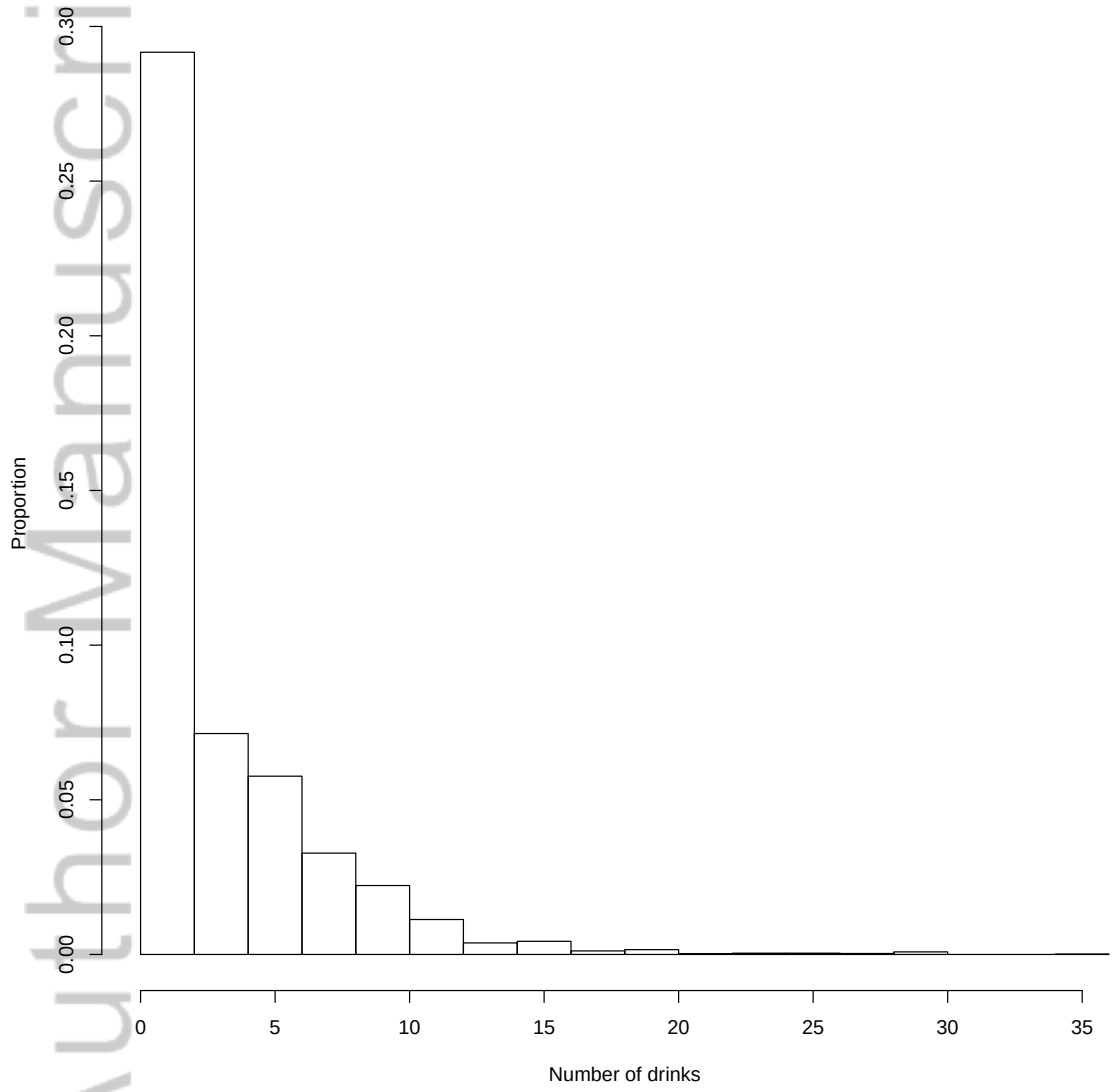

Figure 1. Distribution of the number of drinks per drinking day in past month

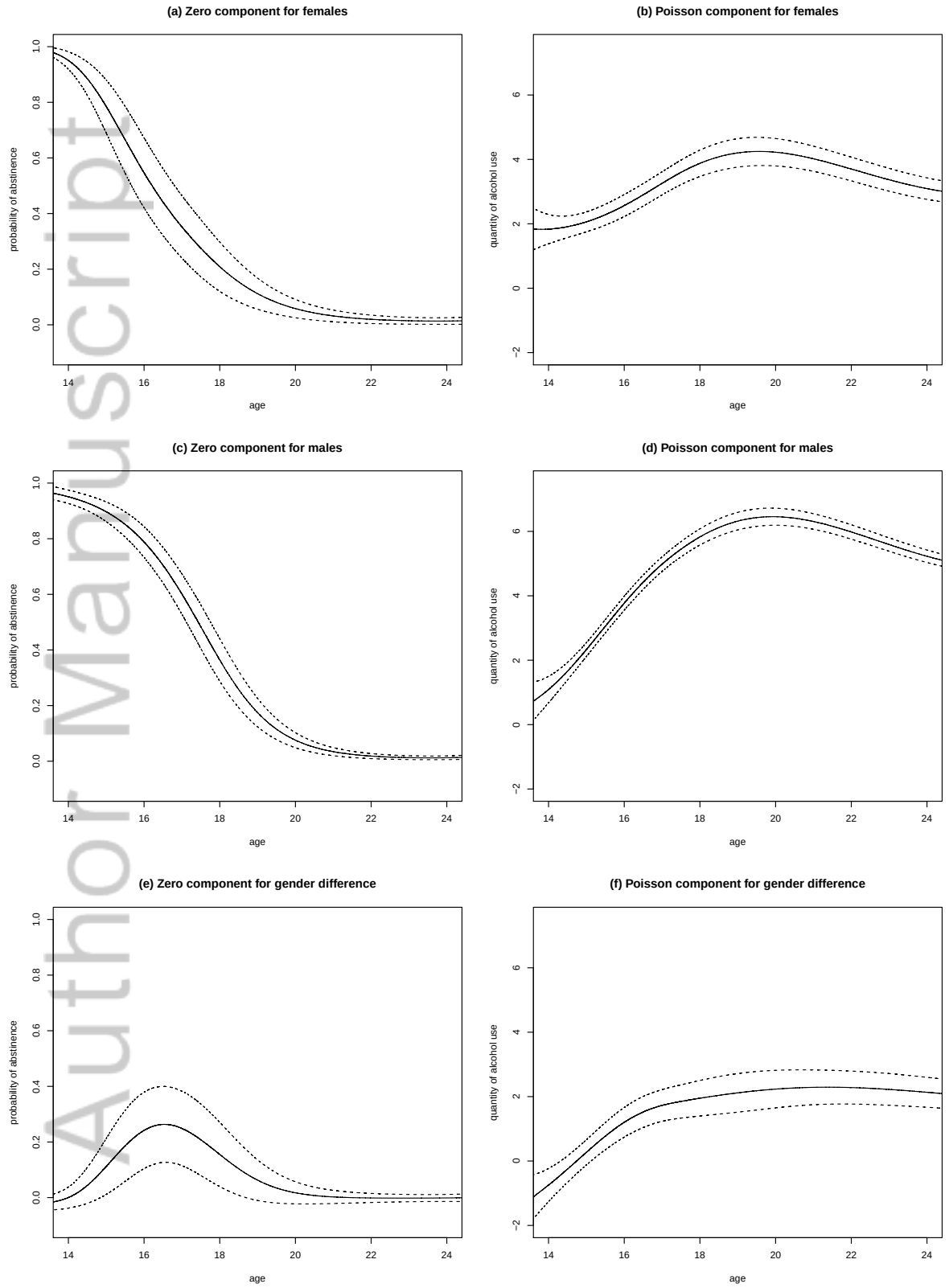

Figure 2. Gender-specific developmental trajectories and time-varying gender differences estimated on the MLS data

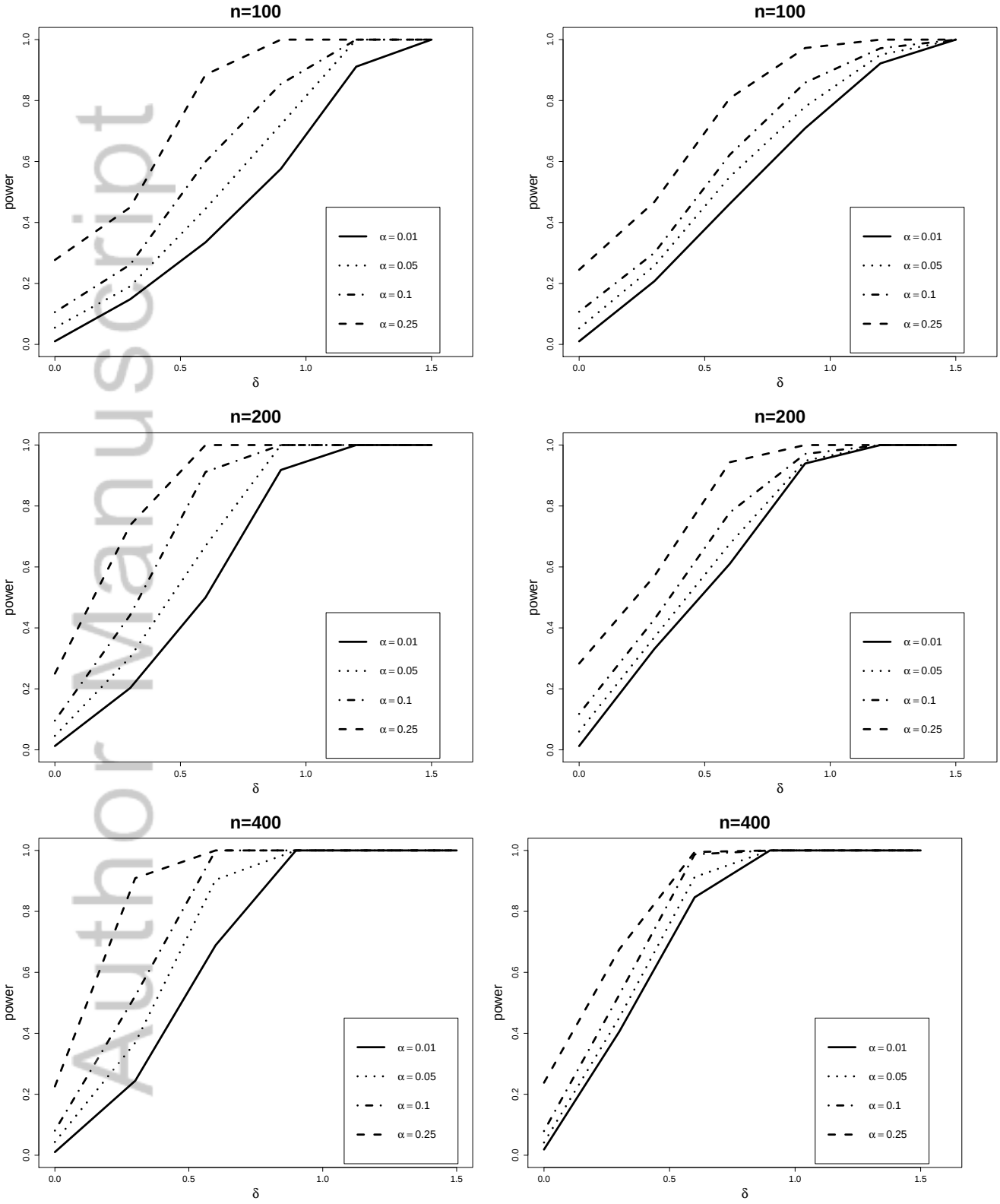

Figure 3. Power curves of the zero component(left) and poisson component(right)