Analytical Methodologies for Clinical Endpoint Data with Excessive Zero

Nancy Liu, Amarjot Kaur, Jing Li, Ziliang Li

Clinical Biostatistics, Merck Research Laboratories, Kenilworth, NJ 07033, U.S.A.

Correspondence to: Nancy Liu, Clinical Biostatistics, Merck Research Laboratories, K15-2/MS 2435, Kenilworth, NJ 07033, U.S.A.
E-mail: nancy_liu@merck.com

Abstract

Zero-inflated normal or log-normal models have been introduced in literature for analyzing continuous data with high frequency of zeros and when the non-zero part of data approximately follow normal or log-normal distributions. However, the focus of available methods has primarily been on the estimations on the parameters of each separate model of the mixed-distribution. This article will discuss the estimation of the overall treatment effect along with its inferences entailing confidence intervals and p-values based on the zero-inflated log-normal model. The inferences based on zero-inflated normal model can be derived similarly. The zero-inflated log-normal model and the analysis of variance method will be compared using simulations and the rescue medication use data from an illustrative example study.

Key words: rescue medication, zero-inflated log-normal, overall treatment effect

1. BACKGROUNDS

Rescue medication use is commonly allowed in clinical trials when evaluating therapies for allergy and other pulmonary diseases. Many a time, the data on use of rescue medication present lots of zero because of lack use of these medications by large number of subjects during the evaluation period of the endpoint. For these reasons, the analysis of such endpoints is challenging because the high frequency of data with zero value, in addition, the non-zero part of the data may not be normally distributed. Zero-inflated log-normal model has been introduced in literature [1] [2] to handle such situations and could be a reasonable candidate or analysis of such data.

The zero-inflated log-normal model consists of two model specifications with a hierarchy. For a random variable Y_i, let R_i represent the occurrence variable where

$$R_i = \begin{cases} 0, & if \;\; Y_i = 0 \\ 1, & if \;\; Y_i > 0 \end{cases} \tag{1}$$

The probability mass function of R is given by

$$P(R_i = r_i) = \begin{cases} 1 - p_i(\theta_1), & if \;\; \mathrm{r}_i = 0 \\ p_i(\theta_1), & if \;\; \mathrm{r}_i = 1 \end{cases} \qquad (2)$$

The occurrence variable is modeled by a logistic model

$$\log it(p_i(\theta_1)) = X_{1i}^{'}\beta_1 \qquad (3)$$

where X_{1i} is a vector of covariance for occurrence. The intensity variable S is defined as the non-zero part of the Y, which is modeled by the log-normal distribution,

$$\log(S_i \mid \theta_2) \sim N(X_{2i}^{'}\beta_2, \sigma_e^2) \qquad (4)$$

where X_{2i} is a vector of covariance for intensity.

The parameters of the occurrence and the intensity variable, $\beta_{1,}$ $\beta_{2,}$ $\sigma_e{}^2$, can be estimated using the maximum likelihood method and the detailed discussion can be found in [1] [2] and the reference therein. However, in order to use this model to analyze endpoints in clinical trials, a single estimate of the overall treatment effect along with its inferences entailing confidence intervals and p-values need to derived based on the distributions of both the occurrence and the intensity variables.

This article will discuss how to estimate the overall treatment effect along with its inferences entailing confidence intervals and p-values. In addition, the zero-inflated log-normal model and the analysis of covariance method will be compared using an illustrative example and simulations.

The rest of this article is organized as follows. Section 2 describes the methods, which include the motivating data from an illustrative example, the zero-inflated log-normal model, and the formulas for the overall treatment effects along with its inferences. Section 3 compares the zero-inflated log-normal model and the analysis of variance model by applying both models to the rescue medication use data. Section 4 examines the properties of the zero-inflated log-normal model by investigating the bias, power and type I error rate through simulations. Some discussions are provided in Section 5.

2. METHODS

2.1 Illustrative Example of Data

Rescue medication use is considered as a clinical endpoint in this example. The medication score for a given day is defined as the sum of scores from all the rescue medications taken based on a given scoring scheme ranging from 0 to 36. The average medication score is formed as the average of the daily score over an observation period.

The histogram of the average medication score over the observation period for the example study is provided in Figure 1. In this example, over 50% of the subjects had a medication score equals to zero. The density of the non-zero part of the score and the density of the non-zero score in log scale are provided in Figure 2 and Figure 3, respectively, which shows the non-zero part of the medication score approximately follows a log-normal distribution.

Based on the distribution of the average score, the zero-inflated log-normal model could be a reasonable candidate for the analysis of such data.

2.2 Model Used For the Example Data

For the outcome variable *Y* (the averages score), we consider the logistic-lognormal mixed-distribution model with treatment and stratum (0 or 1) as covariates for both occurrence and intensity:

$$\begin{aligned} \log it(p_i(\theta_1)) &= a_1 + b_1 z_i + c_1 x_i \\ \log(S_i \mid \theta_2) &\sim N(a_2 + b_2 z_i + c_2 x_i, \sigma_e^2), \end{aligned} \qquad (5)$$

$$E(S \mid \theta_2) = \exp(a_2 + b_2 z + c_2 x + 0.5 \times \sigma_e^2), \qquad (6)$$

Where:

z is the fixed effect treatment, with z = 0 for placebo and z = 1 for active therapy;
X is fixed stratum effect with x = 0 for stratum 1 and x=1 for stratum 2;

The primary interests are the relative treatment difference (the treatment difference relative to placebo effect) and the treatment difference on the average score.

Formula for the Relative Treatment Difference

Consider the ratio of the overall mean for a one unit change in a common covariate Z, i.e.,

$$\frac{E(Y|Z=z+1)-E(Y|Z=z)}{E(Y|Z=z)} = \frac{P(R=1|Z=z+1,\theta_1)}{P(R=1|Z=z,\theta_1)} \times \frac{E(S|Z=z+1,\theta_2)}{E(S|Z=z,\theta_2)} - 1$$

$$= \exp(b_1)\exp(b_2) \times \frac{1+\exp(c_1 x + b_1 z)}{1+\exp(c_1 x + b_1 (z+1))} - 1$$

The treatment difference relative to placebo where z=0 can be calculated as:

$$\exp(b_1)\exp(b_2) \times \frac{1+\exp(c_1 x)}{1+\exp(c_1 x + b_1)} - 1 \qquad (7)$$

Formula for the Treatment Difference

$$\begin{aligned} &E(Y \mid Z = z+1) - E(Y \mid Z = z) \\ &= P(R=1 \mid Z = z+1, \theta_1) \times E(S \mid Z = z+1, \theta_2) - P(R=1 \mid Z = z, \theta_1) \times E(S \mid Z = z, \theta_2) \end{aligned}$$

$$
\begin{aligned}
&= \frac{\exp(a_1 + c_1 x + b_1(z+1))}{1+\exp(a_1 + c_1 x + b_1(z+1))} \times \exp(a_2 + c_2 x + b_2(z+1) + 0.5 \times \sigma_e^2) \\
&- \frac{\exp(a_1 + c_1 x + b_1 z)}{1+\exp(a_1 + c_1 x + b_1 z)} \times \exp(a_2 + c_2 x + b_2 z + 0.5 \times \sigma_e^2)
\end{aligned}
$$

The treatment difference of active therapy and placebo where z=0 can be calculated as:

$$
\begin{aligned}
&\frac{\exp(a_1 + c_1 x + b_1)}{1+\exp(a_1 + c_1 x + b_1)} \times \exp(a_2 + c_2 x + b_2 + 0.5 \times \sigma_e^2) \\
&- \frac{\exp(a_1 + c_1 x)}{1+\exp(a_1 + c_1 x)} \times \exp(a_2 + c_2 x + 0.5 \times \sigma_e^2)
\end{aligned}
\qquad (8)
$$

Statistical Inferences

The estimate of the relative and absolute treatment difference can be derived using the same framework as the Least Squares mean from the regression model. i.e., using the average of fixed effects. The parameters of $\beta_{1,}$ $\beta_{2,}$ $\sigma_e{}^2$, can be estimated using maximum likelihood method; The c_1x and c_2x in the formulas (7) and (8) can be estimated as using the average effect: 0.5*c_1(x=0)+0.5* c_1(x=1)=0.5* c_1, and 0.5* c_2. The average effects weighted by the sample sizes can also be used. For instance, in the example, there was approximately 75% subjects in stratum 1 and 25% subjects in stratum 2, then c_1x and c_2x can be estimated as 0.75* c_1(x=0)+0.25* c_1(x=1)=0.25* c_1, and 0.25* c_2.

The confidence intervals can be derived using the delta method. Although conceptually simple, the calculations of the delta method involved are cumbersome to program in the presence of covariates and are prone to error. With the introduction of SAS PROC NLMIXED, this approach became easily accessible. The estimates of the relative and absolute treatment differences, the delta method-based confidence limits, and the p-values of testing the significance of the relative or absolute treatment difference are output by the procedure, eliminating the need for extensive coding.

The SAS codes for estimations of the relative and absolute treatment effects along with their confidence intervals and p-values based on the zero-inflated log-normal model using the NLMIXED procedure are provided below.

Key Syntax

```
proc nlmixed data=data;
bounds se2>=0;

/* Define the likelihood function of the occurrence */
x1b1=a1+b1*(&var1=1)+c1*(stratum=2);
p=exp(x1b1)/(1+exp(x1b1));
llb=log((1-p)**(1-(Y>0))) + log(p**(Y>0));

/* Define the likelihood of the Intensity and the Y */
x2b2=a2+b2*(&var1=1)+c2*(stratum=2);
pi=arcos(-1);
```

```
E_int=x2b2;

if Y>0 then
  ll=llb+log(1/(sqrt(2*pi*se2)*Y))+(-(log(Y)-x2b2)**2)/(2*se2);
else if Y=0 then ll=llb;


/* Define relative diff */;
r1 = 1 + exp(a1+c1*0.5);
r2 = 1 + exp(a1+c1*0.5+b1);
rc = r1/r2;
reldif = (rc * exp(b1) * exp(b2) - 1)*100 ;

/* Define the treatment difference */
pr1 = exp(a1+ 0.5 * c1)/(1 + exp(a1+ 0.5 * c1));
pr2 = exp(a1+b1+0.5 * c1)/(1 + exp(a1+b1+0.5 * c1));

ee1 = exp(a2+0.5 * c2+se2/2);     *E(S|Z=z);
ee2 = exp(a2+b2+0.5 * c2+se2/2);  *E(S|Z=z+1);

trt_diff = ee2*pr2 - ee1*pr1;

model Y ~ general(ll);

estimate "Relative difference (%)" reldif;
estimate "Treatment difference " trt_diff;

run;
```

3. RESULTS

The average score from the example was analyzed using the zero-inflated log-normal model. The average medication score was also analyzed using the analysis variance model (ANOVA) for comparison. The results are presented in the Table 1. For both relative and absolute treatment differences, the estimates from the zero-inflated log-normal were larger and the p-values were smaller compared to that from the ANOVA model.

Table 1
Comparison of the Zero-inflated log-normal Model with the ANOVA

	Relative Difference		**Treatment difference**	
	95% CI (%)	**P-value**	**95% CI**	**p-value**
ANOVA	-34 (-64, -4)	0.026	-0.6 (-1.3, 0.02)	0.06
Zero-inflated log-normal Model	-49 (-72, -25)	<0.0001	-1.1 (-1.8, -0.3)	0.006

4. SIMULATION

This section compares the performance of the zero-inflated log-normal model with the ANOVA in terms of estimations and statistical inferences of the parameters of interest, such as the relative and the absolute treatment differences.

Data were simulated from a zero-inflated lognormal distribution with:

$$\log it(p_i(\theta_1)) = a_1 + b_1 z_i + c_1 x_i$$
$$\log(S_i \mid \theta_2) \sim N(a_2 + b_2 z_i + c_2 x_i, \sigma_e^2),$$

Different parameters of a_1, b_1, c_1, a_2, b_2, c_2, σ_e^2 are chosen for scenarios listed in Table 2 to Table 5 to study the bias, power, and the type I error of the estimators of the relative and absolute treatment differences. A sample size of 200 per treatment group was used for each dataset; and 5000 replicated datasets were simulated for each scenario.

Table 2
Scenario #1: No effects on either occurrence or intensity (Type I error rate)

	True value	Mean	Standard Error	MSE	Bias	Power (%)	Model
Relative Difference (%)	0.00	2.32	20.90	451.52	2.32	4.02	ANOVA
Relative Difference (%)	0.00	1.26	16.18	254.67	1.26	4.80	Zero-Inflated
Treatment Difference	0.00	0.01	0.61	0.40	0.01	4.58	ANOVA
Treatment Difference	0.00	0.00	0.49	0.23	0.00	4.18	zero-Inflated

Table 3
Scenario #2: Positive treatment effect on the occurrence, no effect on the intensity;

	True Value	Mean	Standard Error	MSE	Bias	Power (%)	Model
Relative Difference (%)	-28.20	-26.29	17.42	315.75	1.92	43.34	ANOVA
Relative Difference (%)	-28.20	-27.03	13.33	187.22	1.18	55.1	Zero-Inflated
Treatment Difference	-0.84	-0.83	0.58	0.36	0.01	33.76	ANOVA
Treatment Difference	-0.84	-0.84	0.46	0.22	0.01	45.3	Zero-Inflated

Table 4
Scenario #3: No effect on the occurrence, Positive effect on the intensity;

	True Value	Mean	Standard Error	MSE	Bias	Power (%)	Model
Relative Difference (%)	-25.92	-23.81	15.84	254.57	2.11	41.74	ANOVA
Relative Difference (%)	-25.92	-24.94	12.00	150.37	0.98	56.78	Zero-Inflated
Treatment Difference	-0.83	-0.81	0.56	0.34	0.03	31.18	ANOVA
Treatment Difference	-0.83	-0.83	0.45	0.21	0.00	44.56	Zero-Inflated

Table 5
Scenario #4: Positive treatment effects on both occurrence and intensity;

	True Value	Mean	Standard Error	MSE	Bias	Power (%)	Model
Relative Difference (%)	-46.81	-45.02	14.03	173.65	1.79	86.36	ANOVA
Relative Difference (%)	-46.81	-45.99	9.86	98.62	0.82	96.42	Zero-Inflated
Treatment Difference	-1.47	-1.44	0.53	0.30	0.03	80.22	ANOVA
Treatment Difference	-1.47	-1.47	0.44	0.19	0.01	93.50	Zero-Inflated

Simulation results show the estimates of the relative and absolute treatment differences using zero-inflated log-normal model appear to be less biased, having smaller variance, more powerful (10%~15% power increase) compared to the estimates using ANOVA. Results also shown the type I error rates were well controlled under 0.05 level.

5. DISCUSSIONS

In this article, we derived the estimation of the overall treatment effects (relative and absolute treatment differences) based on the zero-inflated log-normal model for endpoint with excessive zeros. We also showed how to perform statistical inferences entailing confidence intervals and p-values using the SAS NLMIXED procedure. In addition, the zero-inflated log-normal model and the analysis of variance method were compared using an illustrative example and simulations.

This article showed, when the data approximately follows the zero-inflated log-normal distribution, the estimations of the overall treatment effects derived based on the zero-inflated log-normal model are reliable, accurate, and more powerful compared to the ANOVA model.

The analyses conducted in this article primarily focused on one record per subject type of data. However, the application of such model can be extended to the analysis of data with repeated measures. Tooze et al. [2] proposed a similar model for analyzing longitudinal or repeated measures data with excessive zeros. The estimation of the overall treatment effects (relative and absolute treatment differences) along with their inferences becomes more complicated in the present of correlations within subject, and will be a topic for future research.

REFERENCES

1. Ning Li, David A. Elashoff, Wendie A. Robbins and Lin Xun
   A Hierarchical Zero-inflated log-normal model for Skewed Responses
   Statistical Methods in Medical Research 2008; 00:1-15

2. Janet A Tooze, Gary K Grunwald and Richard H Jones
   Analysis of Repeated Measures Data with Clumping at Zero
   Statistical Methods in Medical Research 2002 11: 341-255

Figure 1:
Histogram of Average Score
Pooled Active Therapy and Placebo

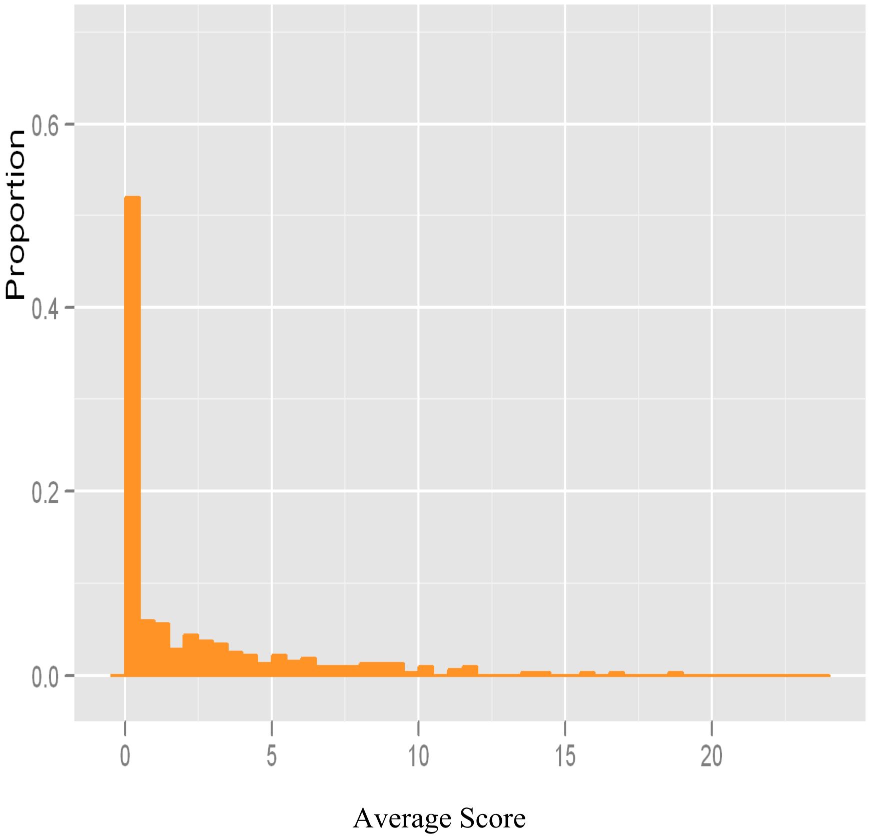


Figure 2
Smoothed Densities of Non-Zero Part of Average Score by Treatment Group

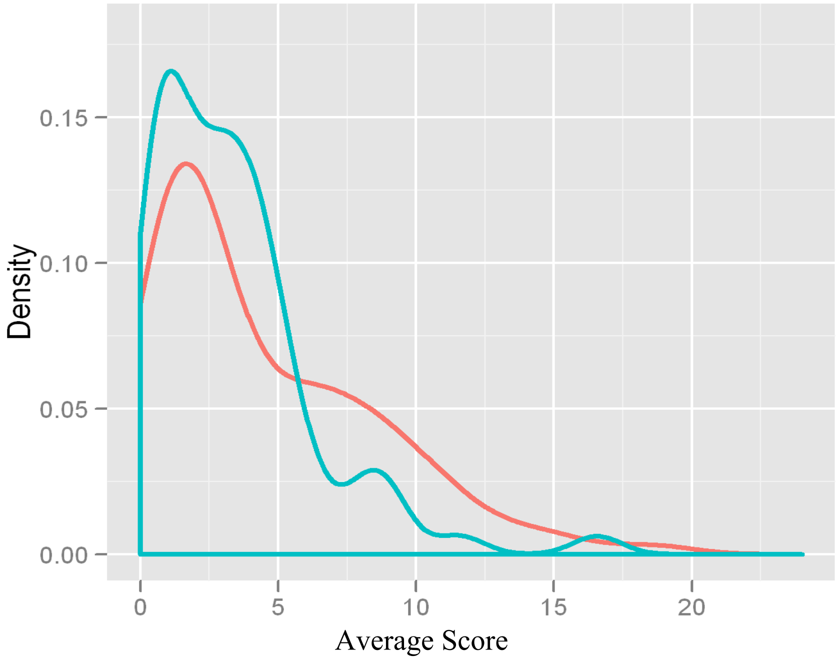
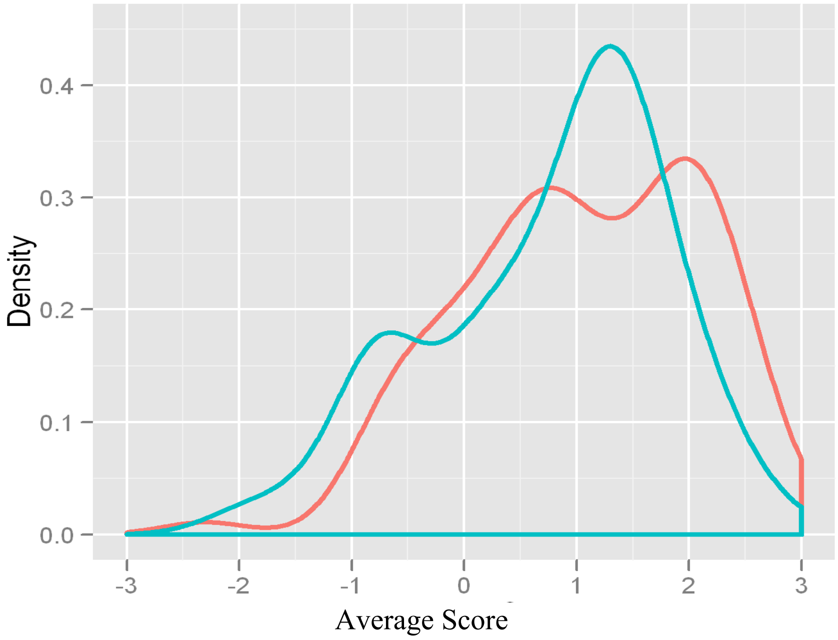


Figure 3
Smoothed Densities of log Non-Zero Part of Average Score by Treatment Group


Density
Average Score