Variance component analysis for proportion data using zero-inflated beta mixed-effects model

Renjie Wei[1]

[1]Novartis Institutes for Biomedical research, Cambridge, MA

ABSTRACT

Variance component (VC) analysis plays an important role in understanding the source of variation observed in the measurement results in a design of experiment. It aims to decompose the total variation observed in the measurement results of an experiment into variation attributable to each experimental factor (i.e. factor-specific variation). Generalized linear mixed-effects models (GLMMs) are commonly used for VC analysis, with different distributions and link functions specified depending on the type of data. However, the distribution of measurements is not always regular. For example, proportions (e.g. DNA allele frequency) measured in many experiments often have an inflated number of zeros. Our simulation studies demonstrated that such zero-inflated proportion data can increase the chance of failure to convergence when GLMMs is applied. Thus, approaches that are more robust to the inflation of zero is needed. Zero-inflated beta (ZIB) mixed-effects model--a two-part model that fit presence/absence of zero by a logistic regression component and non-zero data by a beta regression component—can be considered as an alternative for zero inflated proportion data. ZIB mixed-effects model can decompose variation into two sub-parts: variation in presence of zero, and variation in the non-zero part of data. One need to combine these sub-parts of variation into single estimation for interpretation. We proposed a closed form solution to combine those sub-parts of variation. We also compared the performance of ZIB model with GLMMs. Simulation studies demonstrate that both approaches can give a good estimate when the model convergence is achieved, but ZIB model convergence is more robust to the number of inflated zeros in the data.

INTRODUCTION

- The Oncomine Universal Dx Test targets key regions of 52 unique human genes comprised of approximately 600 somatic variants
- To demonstrate that the test provides reproducible and repeatable results, allele frequency corresponding to each positive DNA variant were measured across multiple sites, operators, instruments and reagent lots
- Estimate variability attributable to each factor

Data and methods

Allele frequency distribution

- Figure 1 shows measured allele frequency distribution for different example variants
- Most variants have approximately normally distributed data (e.g. variant COSM5218 from sample 1)
- A few variants may have excess zeros among data as showed in Figure 1.

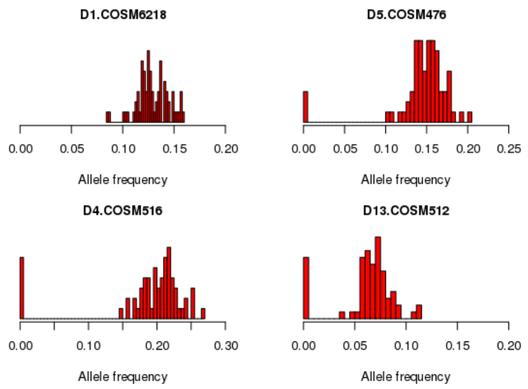

Figure 1: Plots of allele frequency distribution for different variants/samples measured by Oncomine Universal Dx Test

GLMMs based approach for zero inflated proportions

- Gaussian mixed-effects model
- Transformation to shift data from [0,1) to (0,1)
  - y'=[y*(n-1)+0.5]/n
  - Apply Gaussian or Beta mixed-effects model to transformed data y'

ZIB mixed-effects model specification

Assuming y follows a ZIB distribution, the probability density function of y is given by

$$f(y;\mu,\pi,\phi)=\begin{cases}\pi, & y=0\\ (1-\pi)\dfrac{\Gamma(\phi)}{\Gamma(\mu,\phi)\Gamma\big((1-\mu)\phi\big)}, & y\in(0,1)\end{cases}$$

where the mean and variance of y can then be written as

$$E(y\mid\mu,\pi)=z(\mu,\pi)=(1-\pi)\mu$$

$$Var(y\mid\mu,\pi)=h(\mu,\pi)=(1-\pi)\frac{\mu(1-\mu)}{1+\phi}+\pi(1-\pi)\mu^2$$

Add random effects to both the presence/absence of zero, i.e. π, and the mean of positive data, i.e. μ:

$$logit(\mu_i)=\alpha_1+\sum_{j=1}^{m}\gamma_{k_j},\ logit(\pi_i)=\alpha_2+\sum_{j=1}^{m}\delta_{k_j}$$

$$where\ \gamma_{k_j}\sim N(0,\sigma_j)\ and\ \delta_{k_j}\sim N(0,\tau_j)$$

Variance decomposition based on ZIB model

- Decomposition of variance can be made based on the law of total variance.
- The variance between replicate equals to $\mathrm{E}(Var(y_i\mid\mu_i,\pi_i))\approx h(\mu_0,\pi_0)$
- The variance attributable to all m experimental factors equal to $\mathrm{Var}\left(E(y_i\mid\mu_i,\pi_i)\right)=\mathrm{Var}(z(\mu_i,\pi_i))$, which can be approximated as:

$$\left(\frac{\partial z(\mu_0,\pi_0)}{\partial\mu_0}\right)^2\left(\frac{\partial g^{-1}(\alpha_1)}{\partial\alpha_1}\right)^2\sum_{j=1}^{m}\sigma_j+\left(\frac{\partial z(\mu_0,\pi_0)}{\partial\pi_0}\right)^2\left(\frac{\partial g^{-1}(\alpha_2)}{\partial\alpha_2}\right)^2\sum_{j=1}^{m}\tau_j$$

- The formula above demonstrate that variance attributable to the j^{th} factor is a linear combination of variance in presence of zero and variance in positive data, i.e. $\left(\frac{\partial z(\mu_0,\pi_0)}{\partial\mu_0}\right)^2\left(\frac{\partial g^{-1}(\alpha_1)}{\partial\alpha_1}\right)^2\sigma_j+\left(\frac{\partial z(\mu_0,\pi_0)}{\partial\pi_0}\right)^2\left(\frac{\partial g^{-1}(\alpha_2)}{\partial\alpha_2}\right)^2\tau_j$

Simulations and results

Study design

- Data was generated at 4 sites using 3 different reagent lots by 2 operators (nested within site).
- A continuous response measures on $[0,1)$ was generated by a two-stage model approach:
  - Step 1: logistic model to generate zero
  - Step 2: Truncated normal model to generate data on (0,1)
- Four scenarios: 5%, 10%, 15% and 20% zeros
- Four different approach applied:
  - Gaussian model applied to original data
  - Gaussian model applied to transformed data
  - Beta model applied to transformed data
  - ZIB model applied to original data

Results

- Figure 2 shows ZIB model convergence is more robust to excess number of zeros in the data than GLMMs such as Gaussian and Beta mixed models

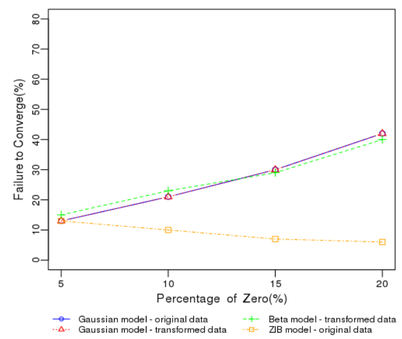

Figure 2: Plots of failure to convergence rate for each approach at different level of zero inflation

- Figure 3 shows both ZIB model and Gaussian/Beta model give good estimates of CV(%) when convergence achieved
- Figure 4 shows ZIB model give good estimate of CV(%) when neither Gaussian nor Beta model converges

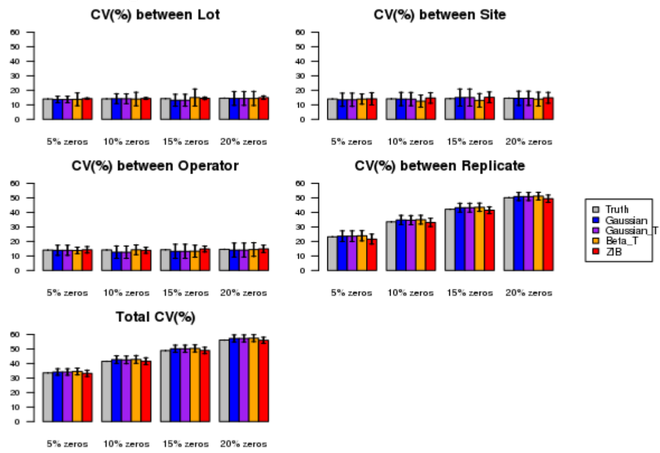

Figure 3: Plots of CV (%) estimates when all approaches achieved convergence

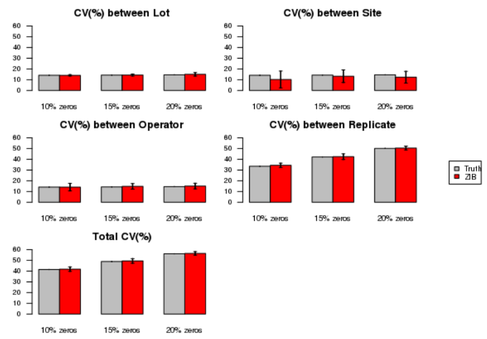

Figure 4: Plots of CV (%) estimates when only ZIB model achieved convergence

Conclusion

This paper proposes an ZIB model based approach for variance component analysis for zero inflated proportion data. Simulation studies demonstrate that our approach is more robust to the excess number of zeros than GLMMs, and can give good estimate when GLMMs fail to converge.

Reference

1. Chen and Li, 2016
2. Ospina and Ferrari, 2010
3. Rigby and Stasinopoulos, 1996
4. Rigby and Stasinopoulos, 2005
5. Rigby and Stasinopoulos, 2007
6. Geweke, 1991

Contact: renjie.wei@novartis.com

NOVARTIS