Randomized Quantile Residuals for Diagnosing Zero-Inflated Generalized Linear Mixed Models with Applications to Microbiome Count Data

Longhai Li

Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, SK, CANADA

31 May 2022
Annual Meeting of Statistical Society of Canada
Virtual Meeting

Acknowledgements

- Thanks to my co-authors of this paper [1]: Bai, W., Dong, M., Li, L., Feng, C., Xu, W., 2021. Randomized quantile residuals for diagnosing zero-inflated generalized linear mixed models with applications to microbiome count data. BMC Bioinformatics 22, 564.

- Thanks to Prof. Pingzhao Hu for inviting me to this session.

- Thanks to NSERC and CFI for providing grants for my research.

Outline

Section 1

Introduction

Introduction

- The next-generation sequencing technologies, such as RNA and microbiome sequencing, typically produce count data measuring the abundance of a large set of nucleic acid sequences. A central goal of analyzing sequencing count data is to identify the sequences with differential abundance under different conditions. For example, the studies in [4] aim to identify microbial taxa with differential abundance in healthy and parkinson patients.
- Generalized linear models (GLM) are commonly used to model the sequencing count data. Negative-binomial (NB) based regression models are used in many widely used bioinformatics analysis tools and methods
- A common drawback of using a parametric model such as a ZINB model is that the model may fail to provide an adequate fit to a dataset. It is challenging to conduct model checking and diagnostics for generalized linear models for count data.

Introduction (contd.)

- The method of randomized quantile residual (RQR) was proposed by Dunn and Smyth [2] to overcome the challenges of diagnosing count regression. *The key idea of the RQR is to randomize the predictive p-value (i.e. tail probability of CDF for response) into a uniform random number.*
- The primary objective of this article is to demonstrate that the method of RQR performs very well for diagnosing zero-inflated GLMMs and is particularly suitable for checking whether such models provide adequate fits to sequencing count data.

Section 2

Zero-inflated/Modified Generalized Linear Mixed Models

Zero-Inflated Models

- A zero-inflated model is a mixture of zero point mass with CDF $F_0(\cdot)$ and a count regression model with CDF $G(\cdot)$. The PMF and CDF of a zero-inflated model can be written as follows:

$$f(y_i) = \begin{cases} p_i + (1-p_i)g(y_i), & \text{for } y_i = 0 \\ (1-p_i)g(y_i), & \text{for } y_i > 0 \end{cases} \quad (1)$$

$$F(y_i = J) = \sum_{j=0}^{J} f(y_i = j) = p_i F_0(J) + (1-p_i)G(J) \quad (2)$$

  where p_i is the mixture proportion.
- In particular, in a ZINB model, the NB distribution is used to model the counts with $g(\cdot)$ given as follows:

$$g(y_i) = f^{\text{NB}}(y_i; \mu_i, \theta) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta)\Gamma(y_i + 1)} \left(\frac{\theta}{\theta + \mu_i}\right)^{\theta} \left(\frac{\mu_i}{\theta + \mu_i}\right)^{y_i} \quad (3)$$

- Typically, p_i and μ_i are linked to covariates through a logistic and log transformation respectively.

Zero-Modified (Hurdle) Models

- In contrast to zero-inflated models, zero-modified models treat zero-count and non-zero outcomes as two separate categories, rather than treating the zero-count outcomes as a mixture of structural and sampling zeros:

$$f(y_i) = \begin{cases} \pi_i, & \text{for } y_i = 0 \\ (1-\pi_i)\frac{g(y_i)}{1-g(0)}, & \text{for } y_i > 0 \end{cases} \quad (4)$$

$$F(y_i = J) = \pi_i F_0(J) + (1-\pi_i)\frac{G(J)-g(0)}{1-g(0)} I(J>0), \quad (5)$$

  where $I(\cdot)$ is the indicator function.
- π_i and μ_i are linked to covariates similarly as in zero-inflated models.

Comparison of Zero-inflated and Zero-modified Models

- When the same $g(\cdot)$ is chosen, the conditional distributions for $y_i|y_i > 0$ in the zero-inflated and zero-modified model are identical—both described with the PMF $g(y_i)/(1 - g(0))$.
- The difference of these two models lies in the modelling of $P(y_i = 0)$.
  - In zero-modified models, $P(y_i = 0) = \pi_i$ is linked to covariates directly.
  - In zero-inflated models, $P(y_i = 0) = p_i + (1 - p_i)g(0)$ is not linked to covariates directly; instead, the mixture proportion p_i is linked to covariates.
- However, we see that when $g(0)$ is very small, which occurs when μ_i is large, these two models are very close.

Section 3

Randomized Quantile Residuals

Definition of Randomized Quantile Residuals

- The method of randomized quantile residual (RQR) [2] was proposed to overcome the difficulties of using traditional residuals for diagnosing regression models for discrete outcomes.
- Let $F(y_i; \mu_i, \phi)$ and $p(y_i; \mu_i, \phi)$ denote the cumulative distribution function (CDF) and probability mass function (PMF) for a model assumed for random variable y_i.
- The randomized tail probability can be defined as:

$$F^*(y_i) = \begin{cases} F(y_i; \mu_i, \phi), & F \text{ is cont. at } y_i \\ F(y_i-; \mu_i, \phi) + u_i\, p(y_i; \mu_i, \phi), & F \text{ is disc. at } y_i \end{cases} \tag{6}$$

where u_i is a uniform random variable on $[0, 1]$, and $F(Y_i-; \mu_i, \phi)$ is the lower limit of F in y_i.

- The RQR for y_i is defined as the normal quantile transformation of $F^*(y_i)$:

$$q_i = \Phi^{-1}(F^*(y_i)). \tag{7}$$

Model Checking with RQR

- Under the true model with **the true parameters**, the distribution of RQRs is a standard normal; see an expository paper [3] by Feng et al. (2020) that explains the normality of RQRs in details through illustrative and simulation studies.
- Based on the normality of RQRs, we can conduct residual diagnostics for count regression models in the same way for normal regression models with Pearson's residuals, including overall GOF tests, graphical examinations such as residual plots and Q-Q plots, and other diagnostics.
- The standard normality holds only when the true model with **the true parameters** is used in Equation (6). The actual performance of the RQR in particular models *with parameters estimated with finite samples* still demands empirical investigation.
- In this paper, we investigate the performance of the RQR in zero-inflated GLMMs with simulated datasets that look like actual microbiome count data.

Section 4

Simulation Studies

Generating Zero-inflated Datasets

The outcome variable Y_i, $i = 1, ..., n$, are generated as follows:

- Generate a binary variable H_i indicating whether Y_i is a structural zero or not, with the probability of $p(H_i = 0) = p_i$, where

$$\log\left(\frac{p_i}{1-p_i}\right) = \tilde{\beta}_0 + \sum_{m=1}^{s} \tilde{\beta}_{X_i^{(m)}} + \sum_{n=1}^{t} \tilde{u}_{Z_i^{(n)}}, \tag{8}$$

  where $\tilde{\beta}_{X_i^{(m)}}$ denotes the mth fixed effect, and $\tilde{u}_{Z_i^{(t)}}$ denotes the tth random effect.

- If $H_i = 0$, $Y_i = 0$; otherwise, Y_i is generated from a NB or Poisson model with mean μ_i:

$$\log(\mu_i) = \log(T_i) + \beta_0 + \sum_{m=1}^{s} \beta_{X_i^{(m)}} + \sum_{n=1}^{t} u_{Z_i^{(n)}}, \tag{9}$$

  where $\beta_{X_i^{(m)}}$ represents the mth fixed-effect, and $u_{Z_i^{(t)}}$ denotes the tth random effect; $\log(T_i)$ denotes the offset term to adjust for the varying total sequence reads.

Generating Zero-modified (Hurdle) Datasets

- π_i is generated similarly as p_i for zero-inflated models, but it represents the proportion of zeros in Y_i rather than the mixture proportion.
- The difference in generating zero-modified datasets is that when $H_i > 0$, Y_i is generated from a truncated Poisson or NB model with mean μ_i.

Parameter settings

- We generate datasets with $s = 3$ fixed factors and $t = 2$ random factors and different sample size $n = 50, 100, 200, 400$.
- The regression coefficients for the fixed-effects covariates $\beta_i \sim N(0, 0.1^2)$ and random effect $u_i \sim N(0, 2^2)$.
- $T_i \sim$ Poisson (300000). The shape parameter θ follows a $2 + unif(0, 1)$ distribution.
- We consider four scenarios by varying $\tilde{\beta}_0$ and β_0:
  - In scenarios 1 and 2, $\tilde{\beta}_0 = 3.5$, which represents the high ZP,
  - In scenarios 3 and 4, $\tilde{\beta}_0 = -5.5$, which represents the low ZP.
  - In scenarios 1 and 3, $\beta_0 = -5.5$ for NB model and $\beta_0 = -5.7$ for Poisson model, which represents the relatively high count data,
  - In scenarios 2 and 4, $\beta_0 = -7.8$ for NB model and $\beta_0 = -8$ for Poisson model, representing the relatively low count data.

Residual Scatterplots and QQ plots for a ZMNB Dataset

Model diagnostics for a single dataset of $n = 400$ samples simulated from a ZMNB model in scenario 4 with parameter settings as low zero proportion and low count.

Figure 1: Graphical Model Checking with RQR and Pearson's Residuals for fitting ZMNB (true model) and ZINB models (close model).

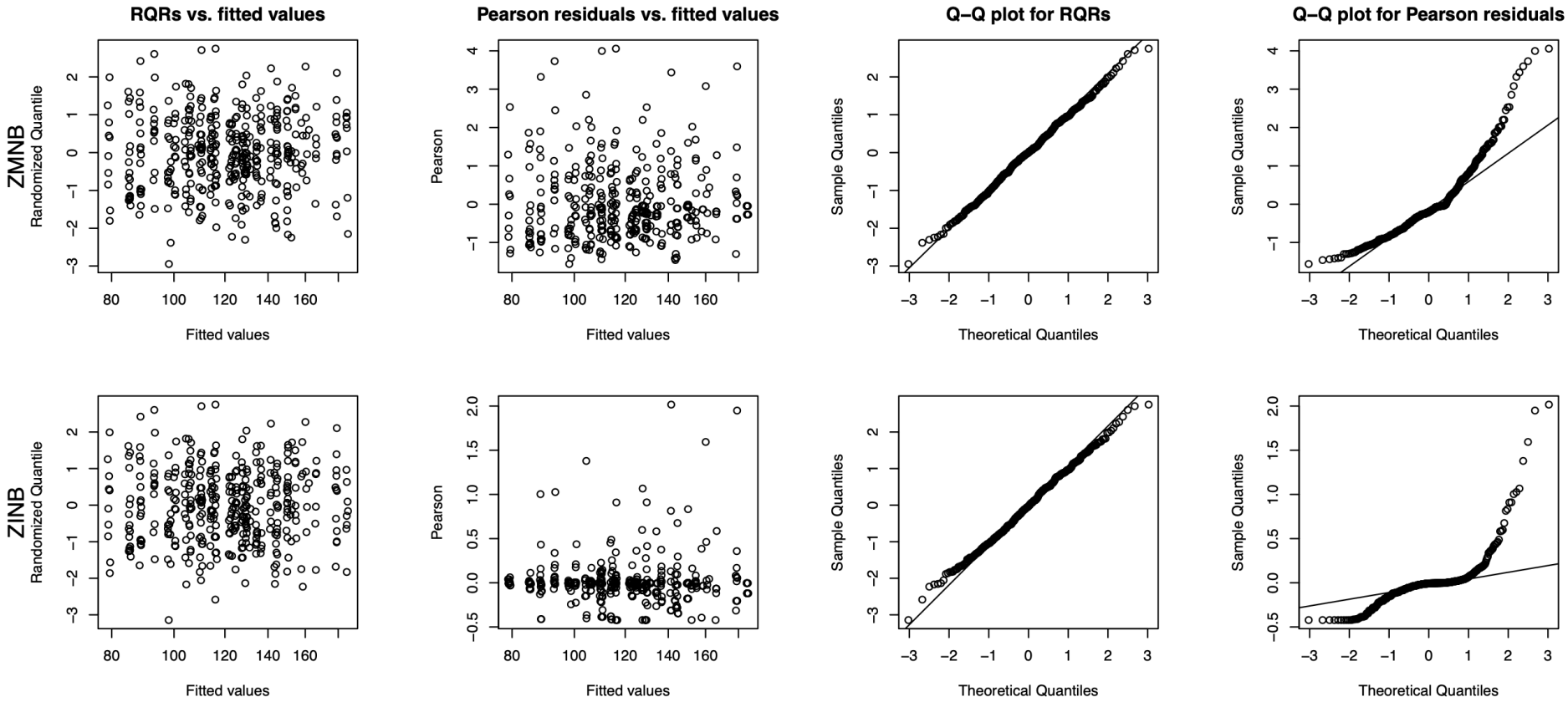
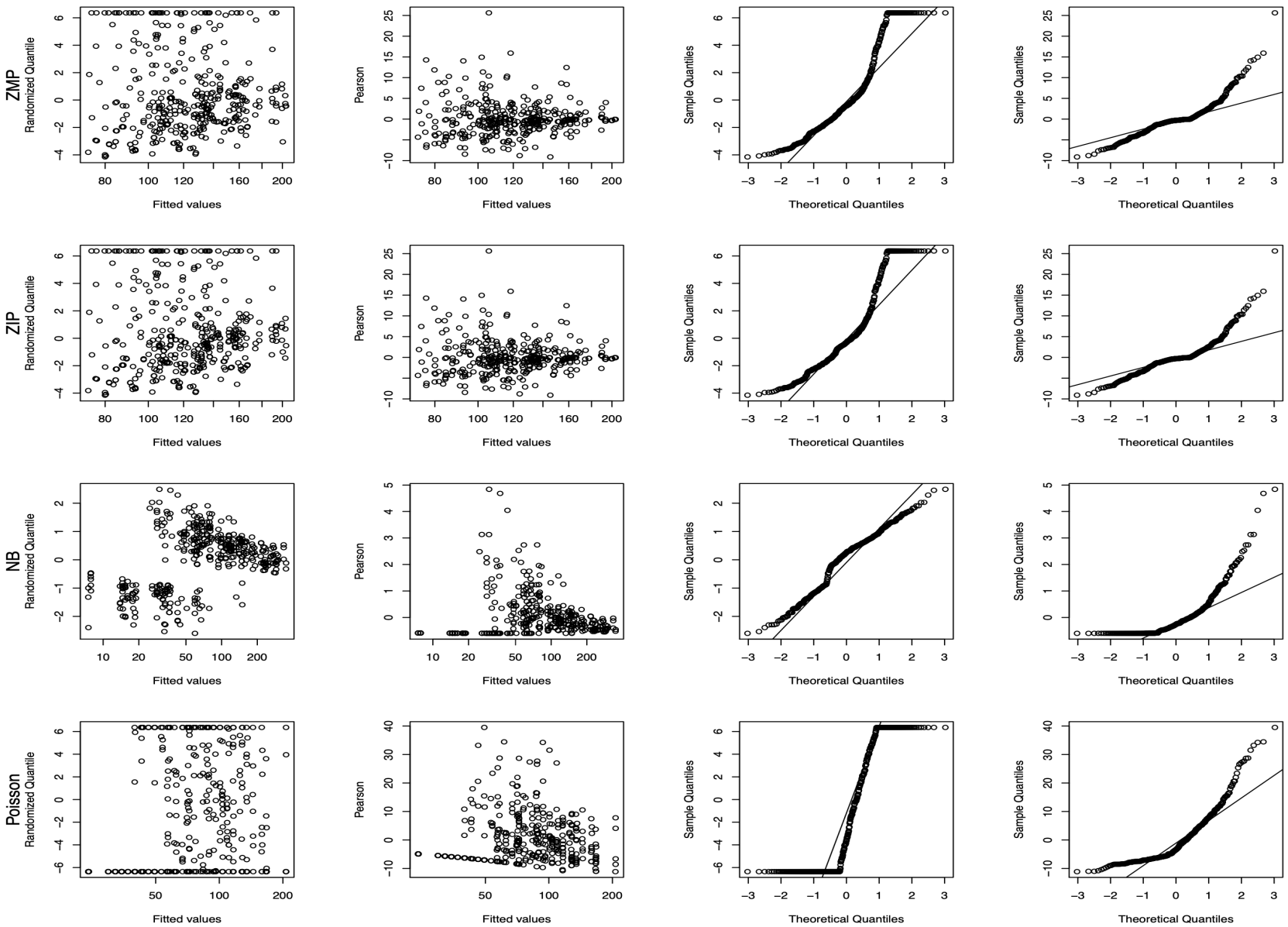


Figure 2: Graphical Model Checking with RQR and Pearson's Residuals for fitting ZMP, ZIP, NB, and Poisson Models (wrong models).



GOF Tests with Multiple Simulated Datasets

The tables in the next pages show the probabilities of rejecting the normality of RQRs based on SW normality test. * represents the true data generating model and † represents the models that theoretically contain or are very close to the true data generating model. ZP is the average zero percentages. The three columns labelled by Q_α show the average of the quantiles of non-zero counts for three α. N is the number of converged model fittings over 3000 replicated datasets.

Table 1: Probabilities of rejecting the normality of RQRs based on SW normality test. Sample size $n = 100$.

Scenario	ZP	$Q_{0.05}$	$Q_{0.5}$	$Q_{0.95}$	ZMNB*	ZINB†	ZMP	ZIP	NB	Poisson	N
1	59	320	1075	2711	0.04	0.05	1	1	0.29	1	1604
2	59	32	109	275	0.04	0.03	1	0.99	0.18	1	1312
3	31	302	1078	2800	0.04	0.04	1	1	0.84	1	1720
4	31	29	107	280	0.05	0.05	1	0.99	0.75	1	1552
Scenario	ZP	$Q_{0.05}$	$Q_{0.5}$	$Q_{0.95}$	ZMNB†	ZINB*	ZMP	ZIP	NB	Poisson	N
1	58	317	1079	2727	0.04	0.04	1	1	0.28	1	1580
2	58	31	108	274	0.04	0.04	1	1	0.17	1	1276
3	31	301	1071	2783	0.05	0.04	1	1	0.85	1	1691
4	31	30	107	279	0.04	0.04	1	1	0.73	1	1535
Scenario	ZP	$Q_{0.05}$	$Q_{0.5}$	$Q_{0.95}$	ZMNB†	ZINB†	ZMP*	ZIP†	NB	Poisson	N
1	55	807	1036	1335	0.03	0.04	0.05	0.04	0.42	1	398
2	56	76	103	138	0.04	0.06	0.06	0.04	0.25	1	339
3	28	768	1008	1320	0.04	0.05	0.05	0.05	0.93	1	633
4	30	73	101	138	0.03	0.04	0.06	0.03	0.83	1	405
Scenario	ZP	$Q_{0.05}$	$Q_{0.5}$	$Q_{0.95}$	ZMNB†	ZINB†	ZMP†	ZIP*	NB	Poisson	N
1	55	793	1020	1318	0.03	0.05	0.03	0.05	0.42	1	390
2	54	74	102	138	0.04	0.04	0.06	0.05	0.30	1	338
3	28	781	1020	1341	0.04	0.03	0.04	0.04	0.92	1	617
4	29	72	102	139	0.03	0.04	0.04	0.04	0.84	1	451

Table 2: Probabilities of rejecting the normality of RQRs based on SW normality test. Sample size $n = 400$.

Scenario	ZP	$Q_{0.05}$	$Q_{0.5}$	$Q_{0.95}$	ZMNB*	ZINB†	ZMP	ZIP	NB	Poisson	N
1	60	285	1068	2808	0.03	0.03	1	1	0.56	1	2475
2	59	28	108	287	0.04	0.04	1	1	0.50	1	2199
3	30	281	1072	2850	0.04	0.05	1	1	0.95	1	2596
4	29	27	107	287	0.04	0.04	1	1	0.90	1	2472
Scenario	ZP	$Q_{0.05}$	$Q_{0.5}$	$Q_{0.95}$	ZMNB†	ZINB*	ZMP	ZIP	NB	Poisson	N
1	60	285	1070	2825	0.04	0.04	1	1	0.57	1	2451
2	60	28	108	286	0.03	0.03	1	1	0.49	1	2212
3	30	280	1069	2843	0.04	0.04	1	1	0.95	1	2613
4	29	27	107	287	0.04	0.04	1	1	0.90	1	2485
Scenario	ZP	$Q_{0.05}$	$Q_{0.5}$	$Q_{0.95}$	ZMNB†	ZINB†	ZMP*	ZIP†	NB	Poisson	N
1	59	777	1012	1315	0.04	0.04	0.05	0.04	0.64	1	906
2	59	74	102	138	0.04	0.03	0.04	0.04	0.57	1	839
3	29	769	1011	1334	0.05	0.05	0.06	0.06	0.97	1	1065
4	29	73	102	139	0.04	0.05	0.06	0.04	0.94	1	960
Scenario	ZP	$Q_{0.05}$	$Q_{0.5}$	$Q_{0.95}$	ZMNB†	ZINB†	ZMP†	ZIP*	NB	Poisson	N
1	59	782	1015	1318	0.04	0.04	0.04	0.05	0.63	1	954
2	59	74	103	139	0.04	0.04	0.04	0.05	0.58	1	816
3	29	769	1015	1340	0.04	0.04	0.06	0.05	0.97	1	1015
4	28	73	102	139	0.04	0.04	0.05	0.05	0.95	1	936

Section 5

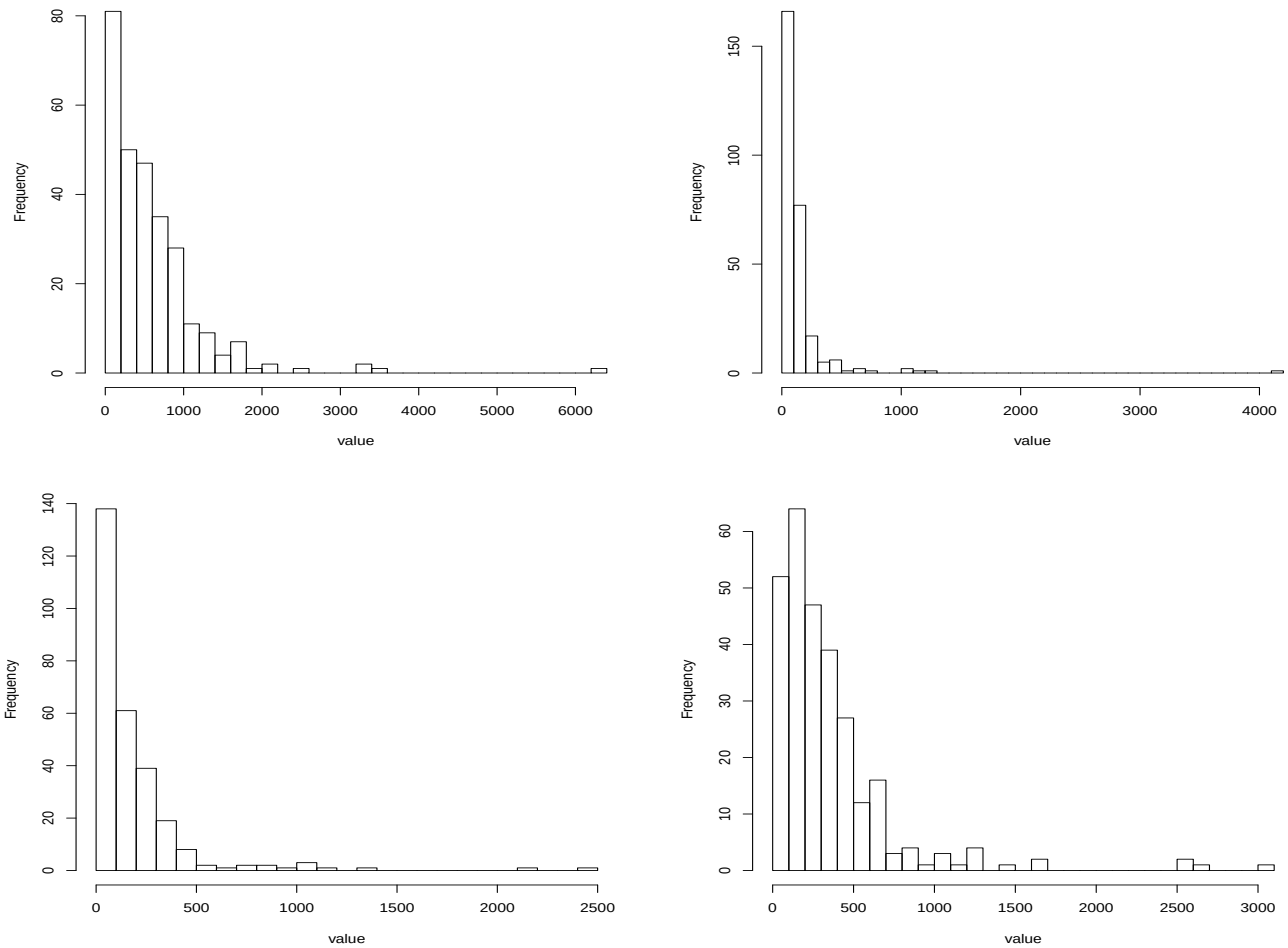
Applications to Check Models for a Microbiome Dataset

Data Sources and Descriptions

- As a response to the epidemic of worldwide obesity, efforts to identify the relationship between host and environmental factors and energy balance have increased.
- The OTU data used in our application were generated at the genus level, which is the commonly used OTU level for microbiome sequencing analysis, and there are 14 different genera in total [5]. Each sample consists of 154 individuals, and we characterize individuals into 31 monozygotic (MZ) twin pairs, 23 dizygotic (DZ) twin pairs and 46 mothers.
- There are 281 OTU measures on the genus level in total. For each measurement, OTU count at each genus level, as well as the total number of reads per measure, were recorded.
- Figure 3 shows the histograms of the four genera selected from the data for the purpose of illustration of the distribution of the OTU measures, which all exhibit right skewness.

Some histograms about the dataset

Figure 3: Histograms of the selected four genera OTUs.

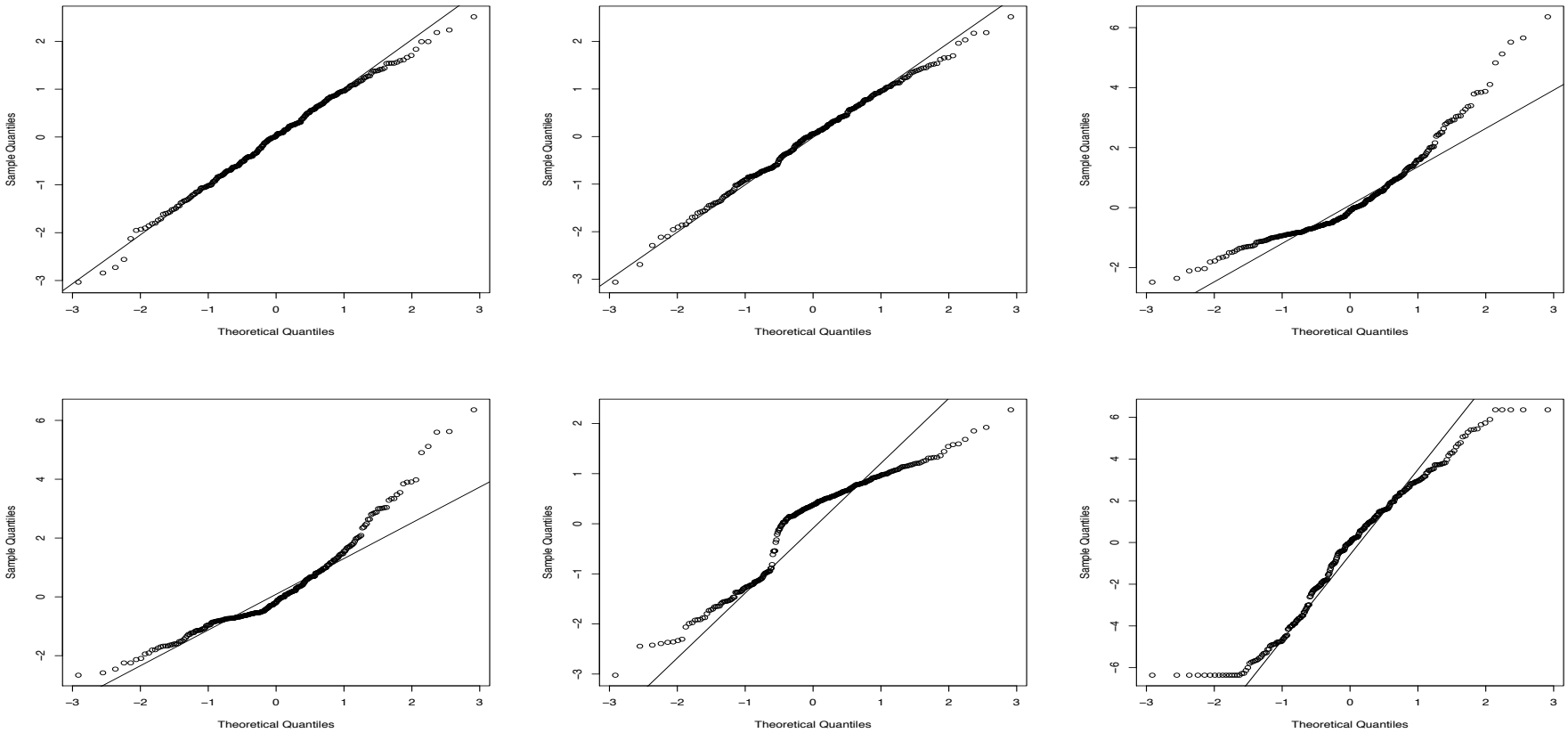


Model Fitting

- In this analysis, ancestry and obesity were selected as the fixed factors while age and family as the random factors.
- Then, ZMNB, ZMP, ZINB, ZIP, NB and Poisson models were fitted to each of the 14 genus-level OTUs. However, the model checking results based on examining the normality of their RQRs showed that these models do not fit the original data very well.
- The OTU counts at the genus level contain very few actual zeros. Therefore, zero-inflated models cannot fit the original data better. Considering that small OTU counts at the genus level are likely caused by the mismatching in sequence alignment of reads, *we truncated the OTU counts to be zero when their values are less than 10 for most genera except the following four genera.* The truncation thresholds for Bacteroides, Ruminococcus,Faecalibacterium and Lachnospiraceae are set to be 50, 50, 100, and 150, respectively. We will present the model diagnostics results for the truncated datasets.

QQ plots of RQR residuals

Figure 4: Q-Q plots of RQRs of six models fitted to Euba OTU data from the Twin Study. The names of models are as follows: ZMNB (top left), ZINB (top middle), ZMP (top right), ZIP (bottom left), NB (bottom middle), Poisson (bottom right).



P-values of the SW normality test applied to RQRs

The rows were sorted according to the p-values of ZMNB models in an ascending order.

Genus	ZMNB	ZINB	ZMP	ZIP	NB	Poisson
Bact	0.052	0.034	$< 10^{-19}$	$< 10^{-19}$	$< 10^{-16}$	$< 10^{-18}$
Lach..g	0.072	0.074	$< 10^{-16}$	$< 10^{-15}$	$< 10^{-3}$	$< 10^{-11}$
Faec	0.083	0.107	$< 10^{-17}$	$< 10^{-18}$	$< 10^{-17}$	$< 10^{-15}$
Rumi	0.232	0.285	$< 10^{-19}$	$< 10^{-19}$	$< 10^{-6}$	$< 10^{-12}$
Rumi.1	0.238	0.366	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-10}$	$< 10^{-11}$
Blau	0.251	0.104	$< 10^{-10}$	$< 10^{-10}$	0.087	$< 10^{-12}$
Erys	0.344	0.258	$< 10^{-16}$	$< 10^{-17}$	$< 10^{-4}$	$< 10^{-7}$
Alis	0.344	0.352	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-9}$	$< 10^{-7}$
Euba	0.461	0.539	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-10}$	$< 10^{-6}$
Lach	0.521	0.358	$< 10^{-9}$	$< 10^{-10}$	$< 10^{-10}$	$< 10^{-5}$
Oscil	0.535	0.606	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-9}$	$< 10^{-5}$
Prev	0.605	0.269	$< 10^{-17}$	$< 10^{-17}$	$< 10^{-4}$	$< 10^{-12}$
Rose	0.627	0.613	$< 10^{-13}$	$< 10^{-14}$	$< 10^{-6}$	$< 10^{-13}$
Copr	0.752	0.721	$< 10^{-13}$	$< 10^{-14}$	$< 10^{-8}$	$< 10^{-6}$

AIC of the competing models for modeling the OTU data

Genus	ZMNB	ZINB	ZMP	ZIP	NB	Poisson
Bact	3698.20	3689.41	29583.11	30276.44	3954.58	52572.41
Lach..g	1096.77	1156.08	Inf	5715.61	1317.49	18248.69
Faec	3328.72	3263.50	13524.32	14132.08	3620.03	29430.38
Rumi	1597.93	1700.08	4495.07	5007.55	1896.48	13263.94
Rumi.1	2432.82	2501.87	6993.35	7536.15	2703.78	13199.43
Blau	3425.68	3401.05	18403.01	18946.44	3396.90	19206.77
Erys	1530.03	1642.23	4129.93	4585.44	1782.82	9082.96
Alis	2159.41	2267.34	4768.40	5300.08	2418.27	9055.12
Euba	2108.20	2123.68	3617.31	4034.78	2292.49	6937.54
Lach	2089.01	2069.09	2325.09	2702.97	2262.49	5286.85
Oscil	1941.61	2044.71	3629.43	4104.53	2218.59	7624.30
Prev	1261.25	1364.03	3757.93	4221.34	1472.21	41117.31
Rose	3234.63	3272.57	18000.67	18547.76	3340.65	21270.47
Copr	2848.91	2829.24	6486.43	6949.32	2914.87	8750.86

Section 6

Conclusions and Discussions

Conclusions

- Our large-scale simulation studies show that the type I error rates of the GOF tests with RQRs are very close to the nominal level. In addition, the scatter plots and Q-Q plots of RQRs are useful in discerning the true and wrong models.
- We also apply the RQRs to diagnose six GLMMs in a real microbiome data analysis. The results show that the OTU counts at the genus level of this dataset after a truncation treatment can be modelled well by zero-inflated and zero-modified NB models.
- R functions for computing RQRs for outputs of R package `glmmTMB` is available in the supplementary file of this paper.

Discussions and Future Work

- The adequate fit of ZMNB and ZINB models to the real microbiome dataset may not be generalized to all microbiome datasets.
- It is of interest to conduct the model diagnostics with RQRs to the ZINB and ZMNB models fitted to a large number of sequencing count datasets.
- In addition to the zero-inflated GLMMs for count data, the RQR method can also be applied to other two-part models, such as zero-inflated beta or zero-inflated log-normal models, for which the randomization needs only to be applied to the observed zeros.

Section 7

References

References

[1] Wei Bai, Mei Dong, Longhai Li, Cindy Feng, and Wei Xu. Randomized quantile residuals for diagnosing zero-inflated generalized linear mixed models with applications to microbiome count data. *BMC Bioinformatics*, 22(1):564, November 2021.

[2] Peter K Dunn and Gordon K Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

[3] Cindy Feng, Longhai Li, and Alireza Sadeghpour. A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, 20(1):175, 2020.

References (contd.)

[4] Erin M. Hill-Burns, Justine W. Debelius, James T. Morton, William T. Wissemann, Matthew R. Lewis, Zachary D. Wallen, Shyamal D. Peddada, Stewart A. Factor, Eric Molho, Cyrus P. Zabetian, Rob Knight, and Haydeh Payami. Parkinson's disease and Parkinson's disease medications have distinct signatures of the gut microbiome. *Movement Disorders*, 32(5):739–749, 2017.

[5] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, January 2009.