# Randomized Quantile Residuals: an Omnibus Model Diagnostic Tool with Unified Reference Distribution

Longhai Li

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

Presented on 19 June 2017
Xiamen University, China

# Acknowledgements

- Joint work with **Alireza Sadeghpour and Cindy X. Feng**.
- The work was supported by grants from Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Foundation for Innovation (CFI).

# Outline

1. Introduction

2. Non-normal Regression Models

3. Pearson and Deviance Residuals

4. Randomized Quantile Residual

5. Simulation Studies
   - Detection of Non-linearity
   - Detection of Zero-Inflation

6. Application to a Big Health Dataset

7. Conclusion and Discussion

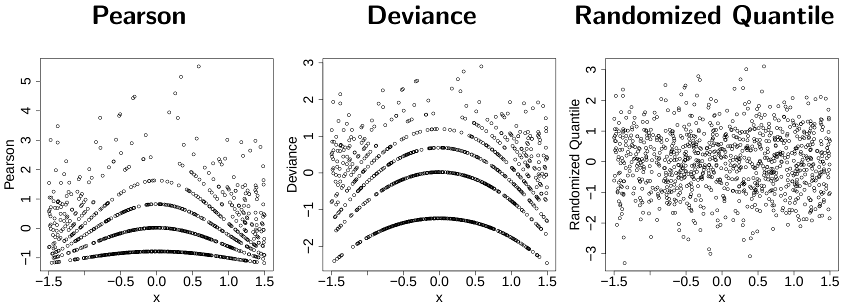
Section 1

# Introduction

# Introduction

- Examining residuals, such as Pearson and deviance residuals, is a primary method to identify the discrepancies between models and data and to assess the overall goodness-of-fit of a model. In normal linear regression, both of these residuals coincide and are normally distributed;
- However, in non-normal regression models, the residuals are far from normality, with residuals aligning nearly parallel curves according to distinct response values, which imposes great challenges for visual inspection.
- Randomized quantile residual was proposed by Dunn and Smyth (1996) to circumvent the above-mentioned problems in traditional residuals.
- Randomized quantile residual is still lack of applications in practice. We will demonstrate how good it is.

# A First Look at Three Residuals



- A simulated dataset is checked against the corrected model. However, Pearson and deviance residuals exhibit trend and cluster in lines.
- In addition, the often used $\chi^2$ tests are not well-calibrated.
- Randomized quantile residuals can be checked as traditional residuals for normal regression.

Section 2

# Non-normal Regression Models

# Generalized Linear Model

- Generalized Linear Model (GLM): The GLM assumes that a response variable $y_i$ given $x_i$ follows a non-normal distribution, such as Poisson and Gamma, etc. A unified form for the PDF is:

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}. \tag{1}$$

A link function is used to connect the conditional expected value of the response variable, $\mu_i = E(y_i|x_i)$, to a linear combination of the covariates and regression parameters as,

$$g(\mu_i) = \eta_i = x_i\beta. \tag{2}$$

# Zero-Inflated Models: I

- Zero-Inflated Model: In practice, very often, we have excessive zeros in count data, which might not be captured by a conventional GLM model. One popular approach to model such data is to use a mixture of point mass at zero modelling the non-risk group (structural zeros) and a GLM modelling the at-risk group.

  **Example**:

  The zero-inflated Poisson response variable with parameters $\lambda_i$ and $p_i$, denoted by $ZIP(\lambda_i, p_i)$, is defined as:

$$y_i \sim \begin{cases} \delta_0 & \text{with probability } p_i \\ \text{Poisson}(\lambda_i) & \text{with probability } 1 - p_i, \end{cases} \tag{3}$$

# Zero-Inflated Models: II

- The PMF of the ZIP distribution:

$$dzip(y_i = 0) = p_i + (1 - p_i)e^{-\lambda_i} \tag{4}$$

$$dzip(y_i = j) = (1 - p_i)\frac{e^{-\lambda_i}\lambda_i^j}{j!},\ j = 1, 2, \cdots . \tag{5}$$

- The CDF of the ZIP distribution

$$pzip(y_i = J; \lambda_i, p_i) = \sum_{j=0}^{J} dzip(y_i = j) = p_i + (1 - p_i)ppois(J, \lambda_i), \tag{6}$$

- Links of $p_i$ and $\lambda_i$ to covariates

$$\text{logit}(p_i) = z_i\gamma \text{ and } \log(\lambda_i) = x_i\beta, \tag{7}$$

where $z_i$ and $x_i$ are vectors of explanatory variables for $p_i$ and $\lambda_i$ with $\gamma$ and $\beta$ corresponding to their parameter vectors, respectively.

Section 3

# Pearson and Deviance Residuals

# Pearson and Deviance Residuals

- *Pearson residual* is defined as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{V}(y_i)}}. \tag{8}$$

- The *deviance residual* for the $i$th observation is defined as signed square root of the corresponding component of $D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}\right)$, i.e.

$$d_i = \mathrm{sign}(y_i - \hat{\mu}_i)\sqrt{2\left\{\omega_i\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\right]\right\}}, \tag{9}$$

where $\tilde{\theta}_i$ and $\hat{\theta}_i$ denote the parameters in the saturated and fitted models, respectively.

# Problems with Traditional Residuals

- In regression models for discrete outcomes, the residuals are far from normality, with residuals aligning nearly parallel curves according to distinct response values, which poses great challenges for visual inspection. Therefore, residual plots for the diagnosis of models for discrete outcome variables give very limited meaningful information for model diagnosis, which renders it of no practical use.
- The *Pearson* $\chi^2$ *statistic* is written as, $X^2 = \sum_{i=1}^{n} r_i^2$, and the *deviance* ($\chi^2$ *statistic)* is written as, $D = \sum_{i=1}^{n} d_i^2$. The asymptotic distribution of $D$ and $X^2$ under the true model is often assumed to be $\chi^2_{n-p}$, where $n$ is the sample size and $p$ is the number of parameters. However, the use of this asymptotic distribution for both $X^2$ and $D$ appears lack of theoretical underpinning.

Section 4

# Randomized Quantile Residual

# Definition of Randomized Quantile Residual

- Predictive p-value for continuous $y_i$:

$$F(y_i; \hat{\mu}_i, \hat{\phi}) = P(Y_i \leq y_i \mid \hat{\mu}_i, \hat{\phi})$$

- Randomized predictive p-value
  If $F$ is discrete, the estimated lower tail probability is randomized into a uniform random number.

$$F^*(y_i; \hat{\mu}_i, \hat{\phi}, u_i) = F(y_i-; \hat{\mu}_i, \hat{\phi}) + u_i \, d(y_i; \hat{\mu}_i, \hat{\phi}), \tag{10}$$

  where $u_i$ from uniform distribution on $(0, 1]$, $F(y_i-; \hat{\mu}_i, \hat{\phi})$ is the lower limit of $F$ at $y_i$, i.e., $\sup_{y<y_i} F(y; \hat{\mu}_i, \hat{\phi})$, the lower limit in the "gap" of $F(\cdot, \hat{\mu}_i, \hat{\phi})$ at $y_i$.

- Randomized quantile residual

$$q_i = \Phi^{-1}(F^*(y_i; \hat{\mu}_i, \hat{\phi}, u_i)) \tag{11}$$

  where $\Phi^{-1}$ is the quantile function of a standard normal distribution

# Illustrative Example #1: I

- **The true model** for $y_i$ has the PMF:

| $y_i$ | 0 | 1 | 2 |
|---|---|---|---|
| $d_0(y_i)$ | 0.25 | 0.5 | 0.25 |

(12)

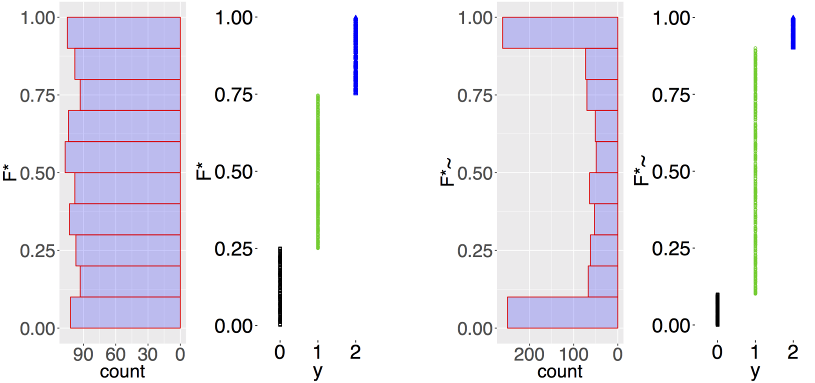
- We compare with a **wrong model** with PMF $d_1$:

| $y_i$ | 0 | 1 | 2 |
|---|---|---|---|
| $d_1(y_i)$ | 0.1 | 0.8 | 0.1 |

(13)

# Illustrative Example #1: II

Figure 2: The randomized predictive p-value for the true model and the wrong model.



(a) $F^*$ for true model

(b) $\tilde{F}^*$ for wrong model

# Illustrative Example #2: I

- **The true model:**
  We simulate a response variable of size $n = 1000$ from a Poisson model with
  $$\log(\mu_i) = -1 + 2\sin(2x_i),$$
  where $\mu_i$ is the expected mean count for the $i$th subject and $x_i \sim Uniform(0, 2\pi)$, $i = 1, \cdots, n$
- **A wrong model:**
  Poisson model with mean structure
  $$\log(\mu_i) = \beta_0 + \beta_1 x_i$$
  with $x_i$ as a predictor with linear effect.

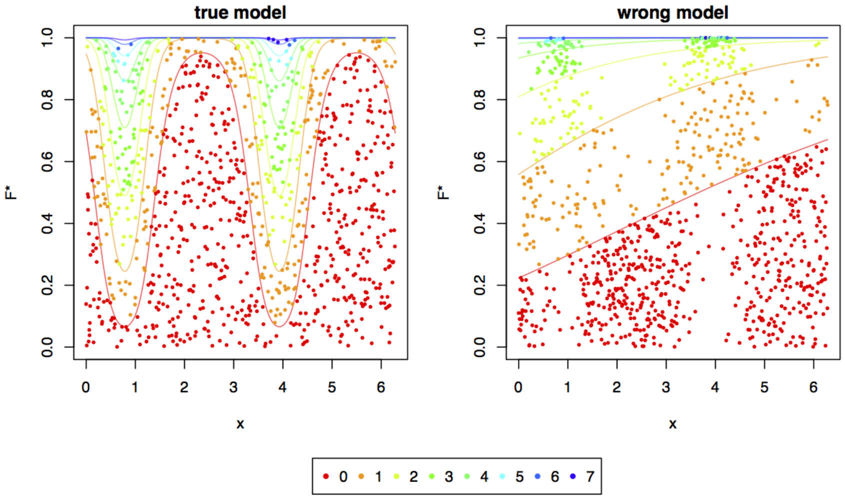
# Illustrative Example #2: II

- CDF lines

  The CDF of the response variable $Y_i$ given $x_i$ (under a considered model with parameters estimated with sample) is denoted by $F(k|x_i) = P(Y_i \leq k|x_i)$, for $k = 0, 1, \cdots$.

- An illustrative picture

# Illustrative Example #2: III



# Normality of Randomized Quantile Residual (RQR)

## Theorem

*Suppose a continuous random variable $Y$ has the CDF $F(y)$, then $F(Y)$ is uniformly distributed on (0,1].*

## Theorem

*Suppose the true distribution of $Y_i$ given $X_i$ has the CDF $F(y_i; \mu_i, \phi)$ and PMF $d(y_i; \mu_i, \phi)$, where $\mu_i$ is a function of $X_i$ involving the model parameters. The randomized lower tail probability $F^*(y_i; \mu_i, \phi, u_i)$ is defined as $F(y_i-; \mu_i, \phi) + u_i\, d(y_i; \mu_i, \phi)$ (10). Suppose $U_i$ is uniformly distributed on (0,1]. Then, we have*

$$F^*(Y_i; \mu_i, \phi, U_i) \sim \mathit{Uniform}(0, 1], \tag{14}$$

*and*

$$q_i = \phi^{-1}(F^*(Y_i; \mu_i, \phi, U_i)) \sim N(0, 1). \tag{15}$$

# Proof of Normality of RQR

For any interval $B \subseteq (0, 1]$,

$$P(F^*(Y_i; \mu_i, \phi, U_i) \in B | Y_i = k^{(j)}) = \frac{\text{length}(F^{(j)} \cap B)}{p^{(j)}},$$

where length$(\cdot)$ is the length of interval. By the law of total probability,

$$\begin{aligned}
& P(F^*(Y_i; \mu_i, \phi, U_i) \in B) && (16)\\
= & \sum_{j=1}^{\infty} P(F^*(Y_i; \mu_i, \phi, U_i) \in B | Y_i = k^{(j)}) \times P(Y_i = k^{(j)}) && (17)\\
= & \sum_{j=1}^{\infty} \frac{\text{length}(F^{(j)} \cap B)}{p^{(j)}} \times p^{(j)} && (18)\\
= & \sum_{j=1}^{\infty} \text{length}(F^{(j)} \cap B) && (19)\\
= & \text{length}(\cup_{j=1}^{\infty} F^{(j)} \cap B) = \text{length}(B) && (20)
\end{aligned}$$

Section 5

# Simulation Studies

Subsection 1

## Detection of Non-linearity

# Simulation Setup

- We simulate a covariate $x \sim \mathit{Uniform}(-1.5, 1.5)$ of size $n = 1000$.
- The response variable is simulated from a **negative binomial regression** model

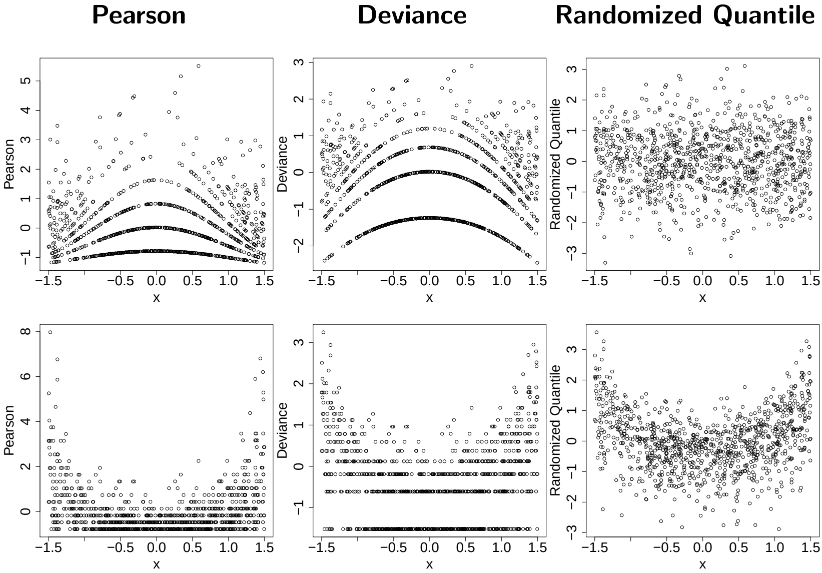
  $$\log(\mu_i) = \beta_0 + \beta_1 x_i^2,$$

  where $\mu_i$ is the expected count for the $i$th subject.
- Then, we consider fitting a **wrong model** assuming

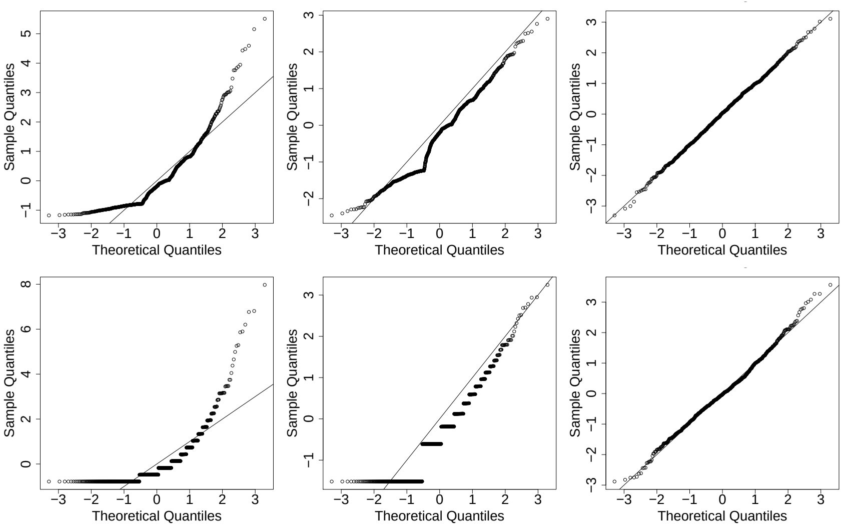
  $$\log(\mu_i) = \beta_0 + \beta_1 x_i.$$

  We set $\beta_0 = 0, \beta_1 = 1$.
- The reciprocal for the dispersion parameter associated with the negative binomial distribution is set as $k = 2$.

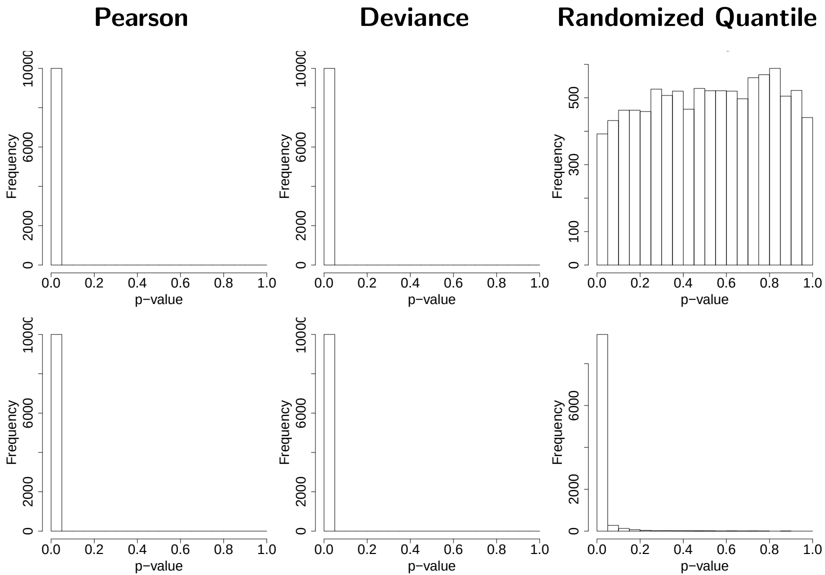
# Scatterplots of Residuals for a Single Dataset



# QQ-plots of Residuals for a Single Dataset



# Replicated Overall GOF Checking with Normality Test

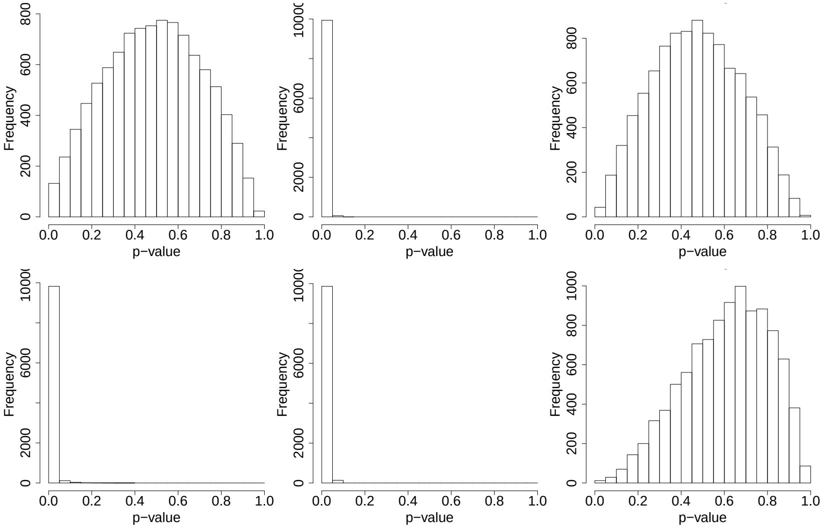


Pearson

Deviance

Randomized Quantile

# Replicated Overall GOF Checking with $\chi^2$ Test



Subsection 2

## Detection of Zero-Inflation

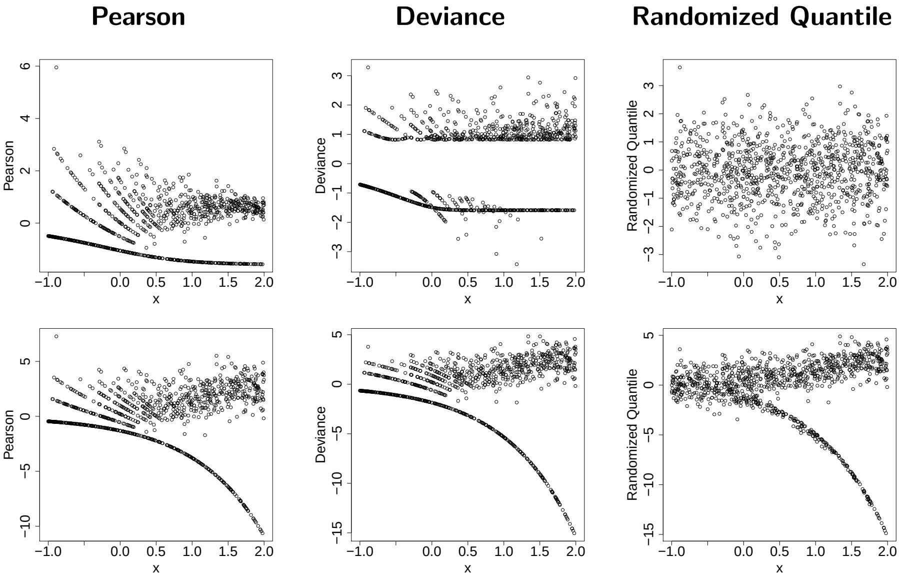
# Simulation Setup

- We simulate a data set of size 1000 with a covariate $x$ from a uniform distribution over $(-1, 2)$ and the response variable from a ZIP model, where the expected mean of the Poisson component is

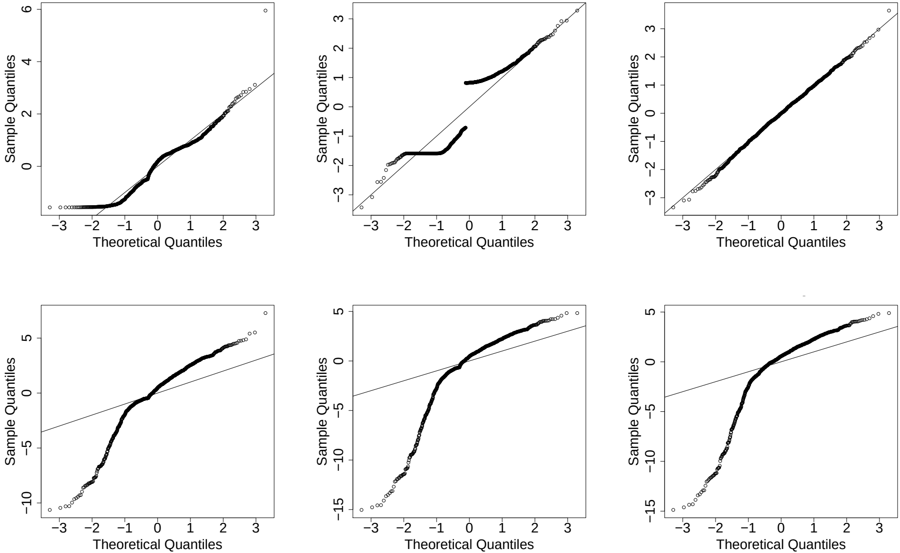
$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

- A Poisson model with the same expected mean $\lambda_i$ is used as an wrong model to compare the power of various residuals.

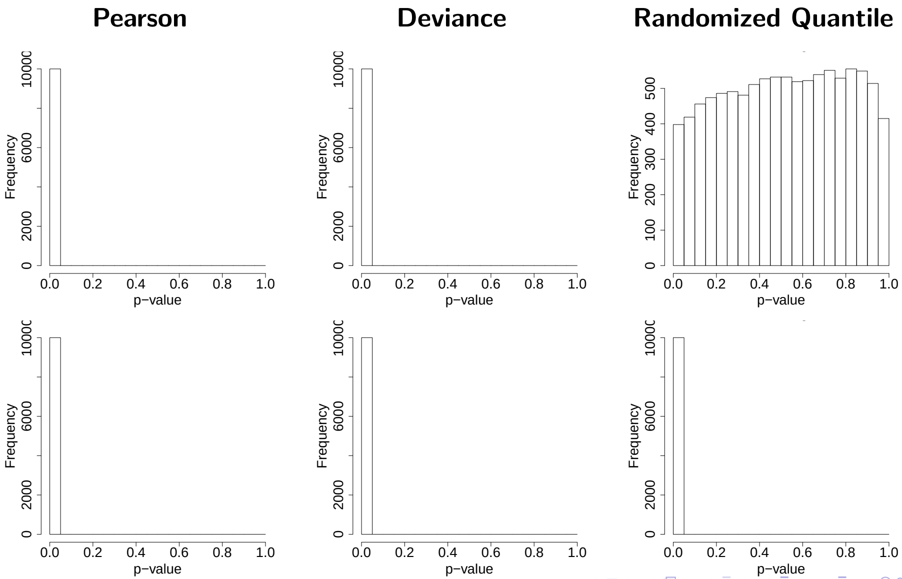
# Scatterplots of Residuals for a Single Dataset



# QQ-plots of Residuals for a Single Dataset



# Replicated Overall GOF Checking with Normality Test

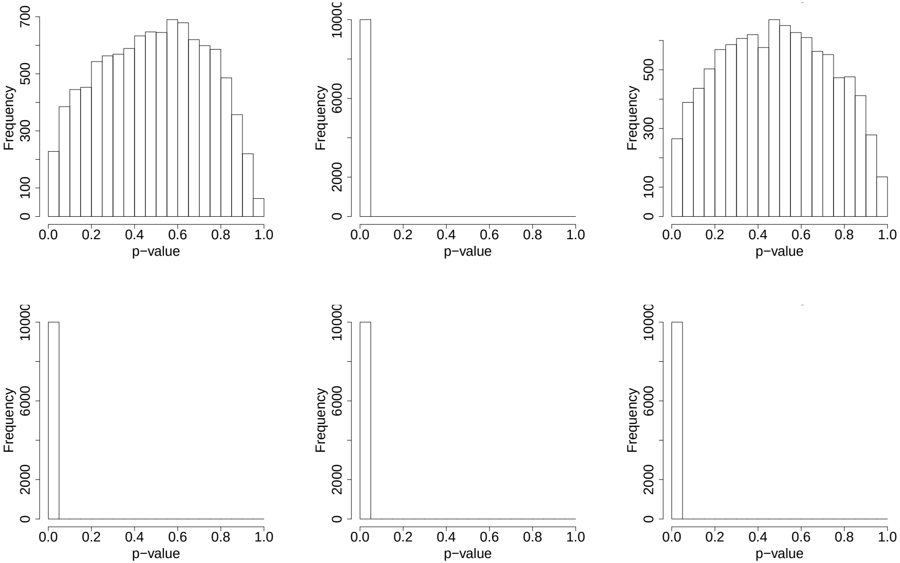


Pearson

Deviance

Randomized Quantile

# Replicated Overall GOF Checking with $\chi^2$ Test



Section 6

# Application to a Big Health Dataset

# Data Description

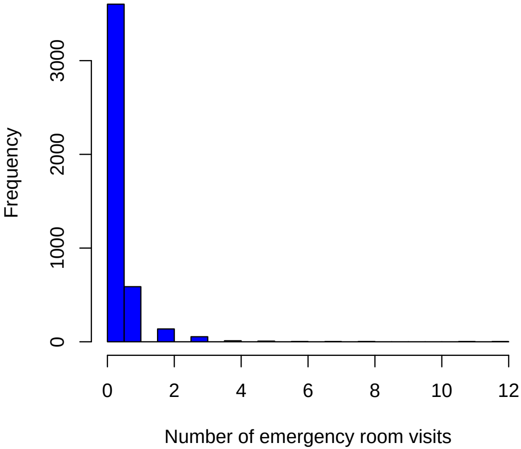
- We apply the randomized quantile residual to examine the goodness-of-fits of the above mentioned more flexible models in a large survey dataset on 4406 individuals, which was collected by the National Medical Expenditure Survey (NMES) for studying the demand of health care of the elderly in the United States.
- The response considered is the number of emergency department visits.
- The covariates include health measures: self-perceived health, the number of chronic conditions, and a measure of disability status; demographic variables: age, race, sex, marital status, education and region; economic variables: family income, employment status, supplementary private insurance status, and public insurance status.

# List of Variables

| Variable | Definition |
|---|---|
| Emer | number of emergency department visits |
| Exclhlth | =1 if self-perceived health is excellent |
| Avghlth | =1 if self-perceived health is average |
| Poorhlth | =1 if self-perceived health is poor |
| Numchron | number of chronic conditions (cancer, heart attack, gall bladder problems, emphysema, arthritis, diabetes, other heart disease) |
| Adldiff | =1 if the person has a condition that limits activities of daily living |
| Noreast | =1 if the person lives in the northeastern US |
| Midwest | =1 if the person lives in the midwestern US |
| West | =1 if the person lives in the western US |
| Age | age in years (divided by 10) |
| Black | =1 if the person is African American |
| Male | =1 if the person is male |
| Married | =1 if the person is married |
| School | number of years of education |
| Faminc | family income in 10,000 |
| Employed | =1 if the person is employed |
| Privins | =1 if the person is covered by private health insurance |
| Medicaid | =1 if the person is covered by Medicaid |

# Histogram of Numbers of Emergency Visits

Figure 12: Frequency distribution for the number of emergency department visits.



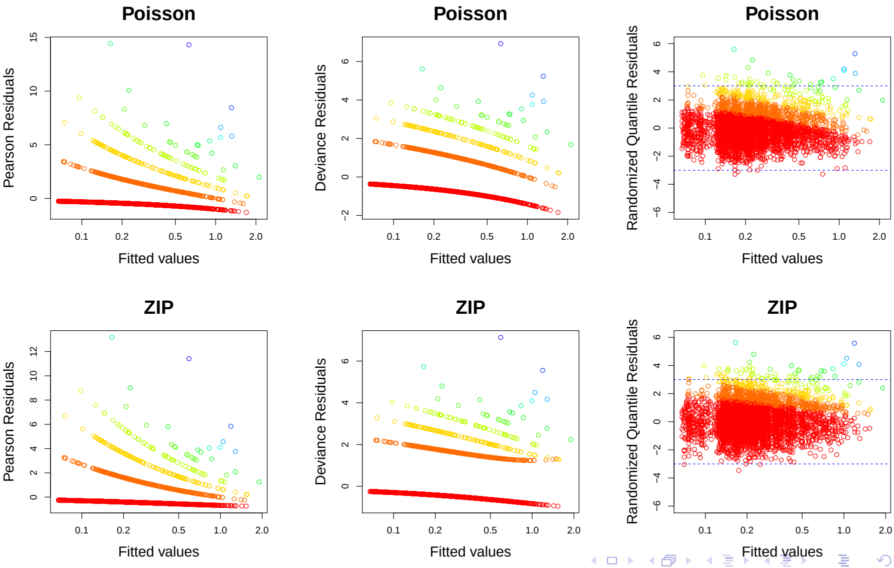
# Four Fitted Models for Further Diagnosis

- We may suspect that there are over-dispersion and/or excessive zero counts; therefore, we consider fitting four models:
  - Poisson
  - Negative Binomial (NB)
  - Zero-inflated Poisson (ZIP) and
  - Zero-inflated Negative Binomial (ZINB)
- Comparing models with Akaike Information Criterion

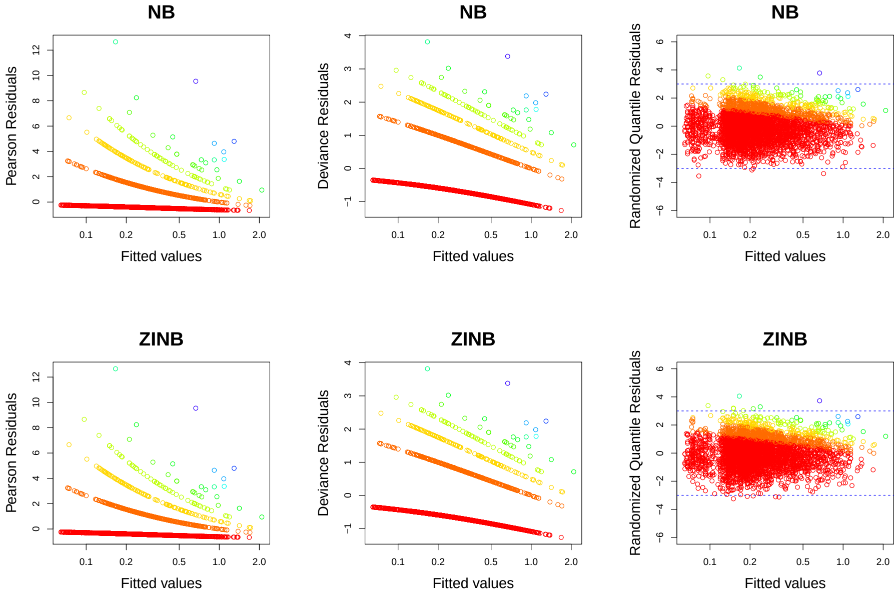
  Poisson=5648, NB=5352, ZIP=5418 and ZINB=5354
- AIC can be used to compare the goodness-of-fit of competing models, but it cannot tell
  - whether a model fits the data adequately,
  - whether the distribution assumption in the response variable is valid.

# Scatterplots of Residuals for Poisson and ZIP

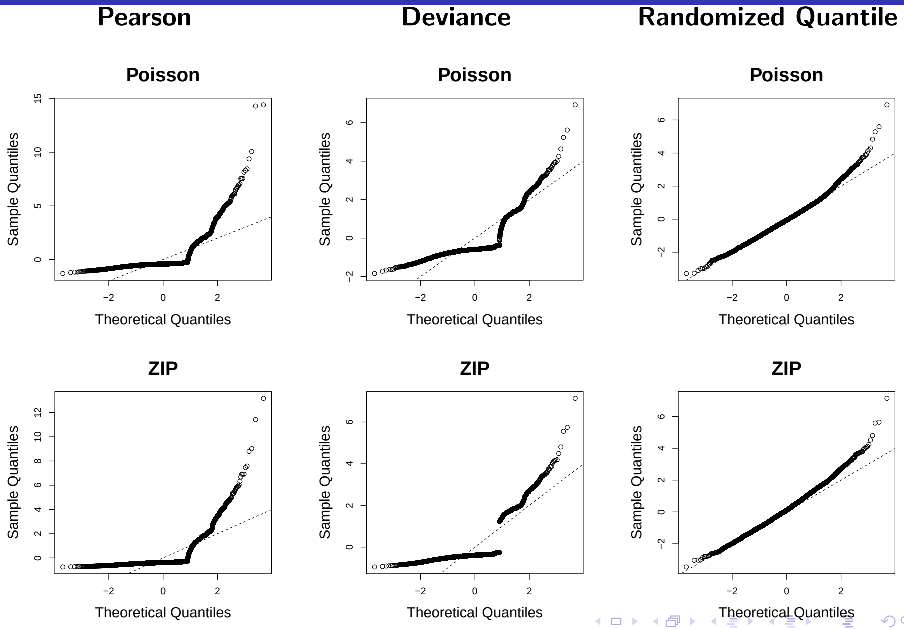
**Pearson** **Deviance** **Randomized Quantile**



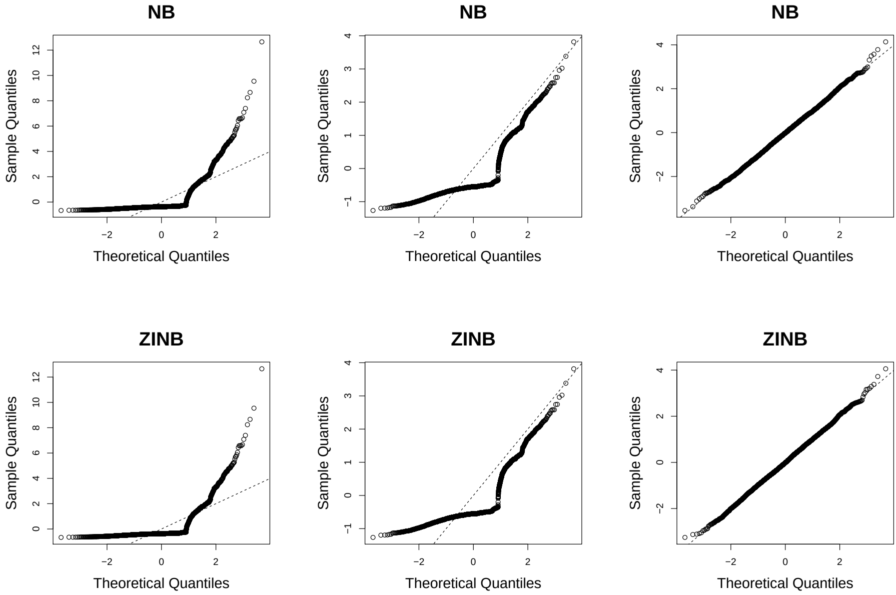
# Scatterplots of Residuals for NB and ZINB



# QQ-plots of Residuals for Poisson and ZIP

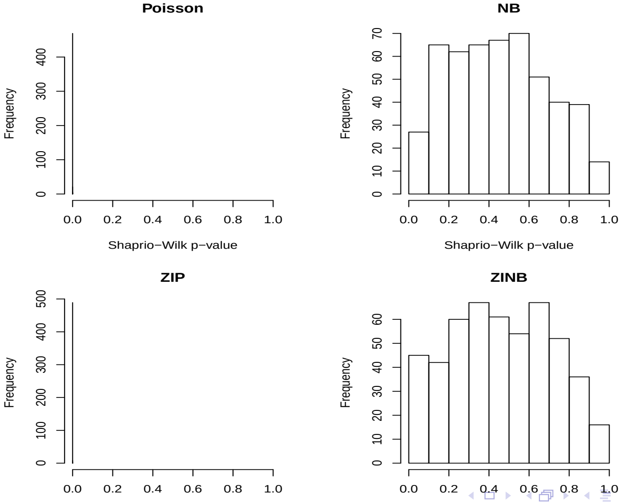


# QQ-plots of Residuals for NB and ZINB



# Replicated Normality Test p-values of RQR

Figure 17: Frequency of the p-values of the Shaprio-Wilk normality test for the 1000 replicated randomized quantile residuals.



Section 7

# Conclusion and Discussion

# Conclusion

Model diagnosis in non-normal regression is very important but difficult. This paper empirically demonstrates that randomized quantile residual is an excellent diagnostic tool: We have empirically demonstrated that

- under any true model, randomized residual quantiles are normal distributed.
- the overall GOF tests by applying normality test to randomized quantile residuals are well-calibrated
- randomized quantile residual has great power in detecting many kinds of model inadequacy

# Future work

- Randomized quantile residual can be applied to mixed-effect models for data with complex structures, such as clustered, temporal, and longitudinal data. We are working to develop residual-based model diagnostic tools for widely used R packages such as `lme4`, and `mgcv` packages in R.
- Randomized quantile residual can be applied to check hierarchical Bayesian models for datasets with complex structure, as an alternative of the widely used posterior predictive checking.
- Utilize randomized quantile residual in many contemporary applications, such as in diagnosing mixed-effect models for high-throughput data.