Robust diagnostics for count regression models

Tsung-Chi Cheng

Department of Statistics
National Chengchi University
Taipei 11605, Taiwan

E-mail: chengt@nccu.edu.tw

Detection of multiple outliers

- Deletion diagnostics
  - single deletion
  - multiple deletion
    - As the number of observations to be deleted increases, there is a combinatorial explosion of the number of deleted subsets to be considered.
- Robust estimator
  - to devise estimators that are not so strongly affected by outliers
- Robust diagnostics
  - a robust model-based technique for the detection of outliers

Outline

- Modeling count variables is a common task in various applications.
  - Poisson regression model
  - Negative binomial (NB) regression model
- Robust estimator
  - Maximum trimming likelihood estimation (MTLE)
  - Minimum trimmed deviances estimator (MTDE)
  - Fast algorithm
- Simulation study
- Real data illustration

Poisson regression model

- The response variable Y, number of occurrences of an event, has a Poisson distribution given k explanatory variables.
- The Poisson regression model is then defined as:

$$f(y_i|\boldsymbol{x}_i) = \exp[-\mu_i + y_i \log \mu_i - \log y_i!], \quad i = 1, 2, \cdots, n,$$

where

$$\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}). \tag{1}$$

- The log of the mean (1) is assumed to be a linear function of the explanatory variables.
- The link function g relates the linear predictor to the expected value μ_i of y_i.
- That is, for $i = 1, 2, \cdots, n$,

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

which is also called as the log-linear model in the context of the generalized linear model (McCullagh and Nelder, 1989).

- See Cameron and Trivedi (1998) and Long and Freese (2006).

Negative binomial regression model

- The classical Poisson regression model for count data is often of limited use in some disciplines because the empirical count data typically exhibit overdispersion and/or an excess number of zeros.
- The negative binomial (NB) regression model is a generalized linear model that accommodates a solution to the overdispersion problem and may function better in the case of excess zeros.
- The negative binomial distribution is defined as:

$$f(y_i|\boldsymbol{x}_i) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}, \quad i = 1, 2, \cdots, n,$$

  where $\Gamma(\cdot)$ is the gamma function and μ_i is corresponding to (1).
- This relaxes the assumption of equality between mean and variance (a property of the Poisson distribution) since the variance of the negative binomial distribution is equal to $\mu + \alpha\mu^2$, where $\alpha \geq 0$ is a dispersion parameter.
  - If $\alpha = 0$, then the negative binomial distribution reduces to Poisson.
- Lawless (1987) discusses the statistical properties of negative regression model (e.g. Hilbe 2011).

MLE and MDE

- These parameters can be estimated using the method of MLE by maximizing

$$L(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} l(\mu_i; y_i), \tag{2}$$

  - $l(\mu(\boldsymbol{x}_i^T \boldsymbol{\beta}); y_i)$ denotes the log likelihood for the ith case.
- As the log likelihood function is defined up to an additive constant by $l(\mu, y)$, the deviance function is defined as:

$$d(\mu_i; y_i) = -2\{l(\mu_i; y_i) - l_{\max}(\mu_i; y_i)\}$$

  - $l_{\max}(\mu_i; y_i)$ is the maximum of $l(\mu_i; y_i)$ with respect to μ_i.
- Thus, maximizing $L(\boldsymbol{\mu}; \boldsymbol{y})$ in (2) for MLE, $\hat{\boldsymbol{\beta}}$, is formally equivalent to minimizing (Pregibon, 1982)

$$D(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} d(\mu(\boldsymbol{x}_i^T \boldsymbol{\beta}); y_i). \tag{3}$$

  - The MLE, $\hat{\boldsymbol{\beta}}$, satisfies the minimization of (3).
  - Pregibon (1982) shows how the MLE is related to the minimum deviance estimation (MDE) for a logistic regression model.

Deviance residual for NB

- The log-likelihood for the ith observation

$$\begin{aligned}\ell_i &= \ell(\boldsymbol{\beta}, \alpha; y_i) = \log(\Gamma(\alpha^{-1} + y_i) - \log(\Gamma(y_i + 1)) - \\ &\quad \log(\Gamma(\alpha^{-1})) + y_i \log(\alpha\mu_i/(1 + \alpha\mu_i)) - \alpha^{-1} \log(1 + \alpha\mu_i)\end{aligned}$$

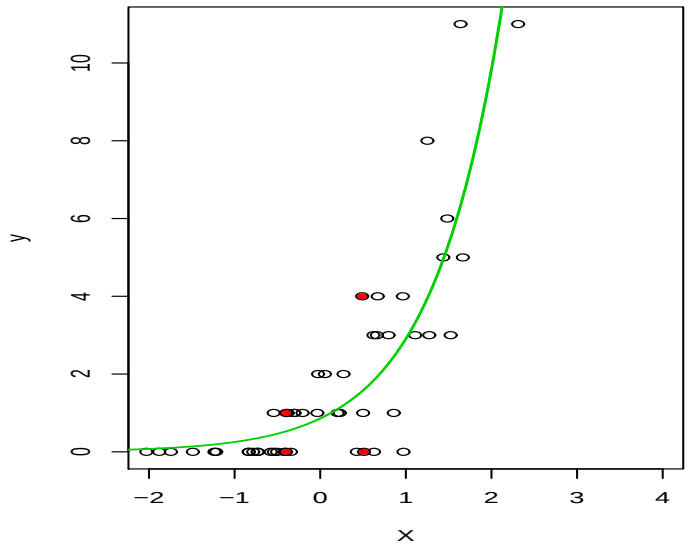
- Once the estimate of $\boldsymbol{\beta}$ is obtained, denoted by $\hat{\boldsymbol{\beta}}$, the estimated value of the model is $\log(\hat{\mu}_i) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$.
- The signed square-root deviance residual for the ith observation is then

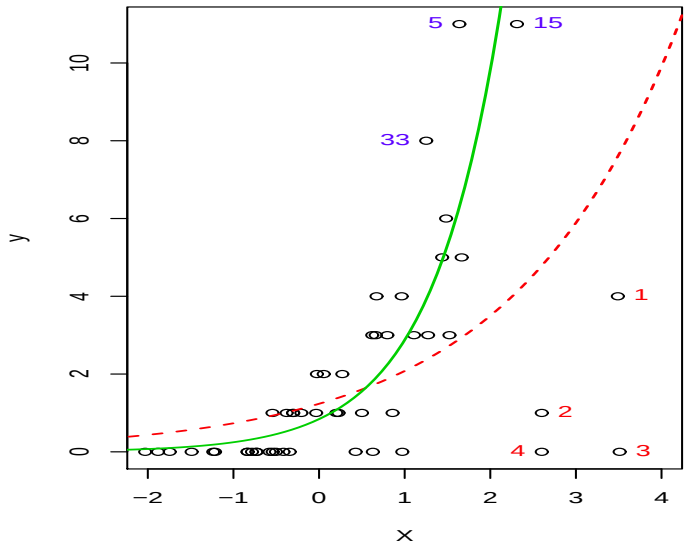
$$\begin{aligned}d_i^* &= d(y_i; \hat{\mu}_i, \hat{\alpha}) = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2} \left\{ \frac{1}{\hat{\alpha}} \log \frac{1 + \hat{\alpha}\hat{\mu}_i}{1 + \hat{\alpha} y_i} \right. \\ &\quad \left. + y_i \log \frac{y_i(1 + \hat{\alpha}\hat{\mu}_i)}{\hat{\mu}_i(1 + \hat{\alpha} y_i)} \right\}^{1/2} \qquad (4)\end{aligned}$$

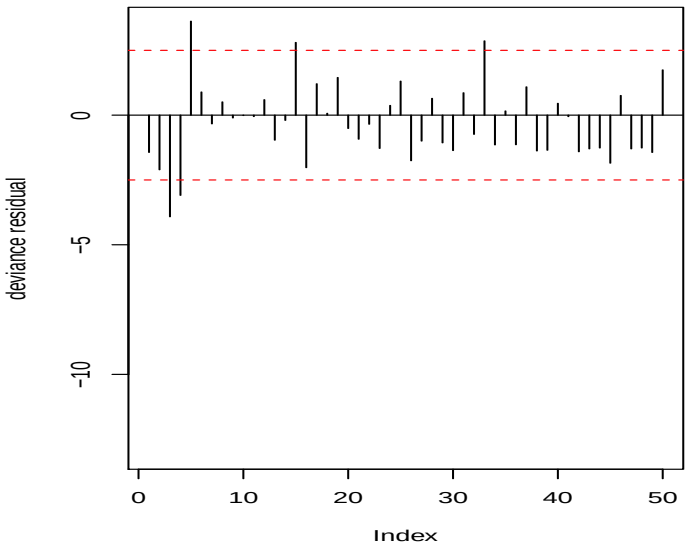
Outliers in count regression model

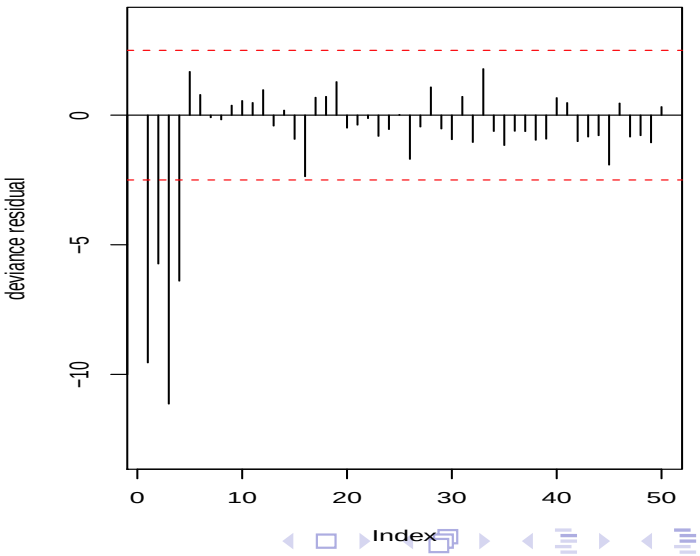
- Classical diagnostics (Pregibon 1981)
- Influence diagnostics for NB regression (Svetliza and Paula, 2001 and 2003)
- Two effects due to multiple outliers
  - Masking effect
    - Outlying cases appear inlying
  - Swamping effect
    - Inlying cases appears outlying

Synthetic data









Breakdown point

- Donoho and Huber (1983)
- Consider all possible corrupted samples Z' that are obtained by replacing any m of the original data points Z by arbitrary values.
- The maximum bias that can be caused by such a contamination is

$$\text{bias}(m; T, Z) = \sup_{Z'} ||T(Z') - T(Z)||,$$

  where the supremum is over all possible Z' and $T(Z)$ is an estimator.
- The (finite sample) breakdown point of estimator T at the sample Z is defined as

$$\epsilon_n^*(T, Z) = \min\left\{\frac{m}{n}; \text{bias}(m; T, Z) \text{ is finite}\right\}.$$

  - It is the smallest fraction of contamination that can cause the estimator T to take on values arbitrarily far from $T(Z)$.
  - For OLS regression one unusual case is enough to influence the coefficient estimates. Its breakdown point is then $1/n$.

Robust generalized linear regression model

- In the presence of multiple outliers, inferences based on the MLE (or MDE) for the generalized linear model are generally inconsistent, while robust estimators support reliable inferences (e.g. Kunsch, Stefanski, and Carroll 1989; Andersen 1992; Cantoni and Ronchetti 2001; Mebane and Sekhon 2004).
- There have been attempts to explore and develop robust methods for the generalized linear models in the literature.
  - For example, Stefanski *et al.* (1986), Künsch *et al.* (1989), Christmann (1994), Atkinson and Riani (2001), Christmann and Rousseeuw (2001), Rousseeuw and Christmann (2003), Müller and Neykov (2003), Čížek (2008), Bianco and Martínez (2009) and Bergesio and Yohai (2011).
  - Valodra and Yohai (2014)
- R functions
  - glmrob in robustbase package
  - fwdglm in forward package

Maximum trimmed likelihood estimator (I)

- Proposed by Hadi and Luceño (1997) and Vandev and Neykov (1998).
- A trimmed likelihood principle is based on trimming the likelihood function rather than directly trimming the data.
  - It is always possible to order and trim observations according to their contributions to the likelihood function, because the likelihood is scalar-valued.
- The trimmed likelihood approach considers to maximize the following objective function:

$$\sum_{i=a}^{b} w_{(i),\nu} l(\theta; y_{(i),\nu}), \tag{5}$$

where $a \leq b,\ \nu = (a, b) \in \{1, 2, \cdots, n\}$, and

$$l(\theta; y_{(1),\nu}) \geq l(\theta; y_{(2),\nu}) \geq \cdots \geq l(\theta; y_{(n),\nu}), \tag{6}$$

for any given value of θ.

- Here $l(\theta; y_i) = \log f(y_i; \theta)$ is the contribution of the ith observation to the log likelihood function.

Maximum trimmed likelihood estimator (II)

- Neykov *et al.* (2007) give the combinatorial representation of MTLE as follows:

$$\begin{aligned}\max_{\theta} \sum_{i=1}^{q} w_{(i),\mathcal{Q}} l(\theta; y_{(i),\mathcal{Q}}) &= \max_{\theta} \max_{\mathcal{Q} \in Q} \sum_{i \in \mathcal{Q}} w_i l(\theta; y_i) \\ &= \max_{\mathcal{Q} \in Q} \max_{\theta} \sum_{i \in \mathcal{Q}} w_i l(\theta; y_i),\end{aligned}$$

  where Q is the set of all q-subsets of the set $\{1, \cdots, n\}$.

- It follows that all possible $\binom{n}{q}$ partitions of the data have to be fitted by the MLE, and the MTLE is given by the partition with the maximum log-likelihood.

Minimum trimmed deviances estimator

- Let $\boldsymbol{\beta}_q$ be the parameter vector for a specific value of q and $\mathcal{Q}$ be the subset with those q cases.
- The corresponding data of $\mathcal{Q}$ are denoted by $\boldsymbol{y}_q$ and $\boldsymbol{X}_q$.
- Consider the ordered sequence

$$d_{(1),\mathcal{Q}} \leq d_{(2),\mathcal{Q}} \leq \cdots \leq d_{(n),\mathcal{Q}}, \tag{7}$$

  where $d_{(i),\mathcal{Q}}$ denotes the ith-ordered deviance residual and $d_{i,\mathcal{Q}} = d(\mu(\boldsymbol{x}_i^T \boldsymbol{\beta}_q); y_i)$.
- The MTD estimator evaluated at q is derived by taking:

$$\min_{\beta_q} \sum_{i \in \mathcal{Q}} d_{(i)}(\mu(\boldsymbol{x}_i^T \boldsymbol{\beta}_q); y_i), \tag{8}$$

  which is denoted by $\hat{\boldsymbol{\beta}}_q$.
- Under this restriction, the MTD estimator corresponds to the MDE and MLE based on the subset $\mathcal{Q}$.
- The corresponding signed square-root deviance residuals in (4) will be modified by replacing the estimated model, $\log(\hat{\mu}_i) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_q$, for $i = 1, 2, \ldots, n$.

MTDE

- Consider Q to be the set of all q-subsets of the set $\{1, ..., n\}$, adapting the combinatorial representation for MTLE of Neykov *et al.* (2007).
- The MTD estimator evaluated at q is shown as follows:

$$\begin{aligned} \min_{\beta} \sum_{i=1}^{q} d^2(\mu_{(i),\mathcal{Q}}; y_{(i),\mathcal{Q}}) &= \min_{\beta} \min_{\mathcal{Q} \in Q} \sum_{i \in \mathcal{Q}} d(\mu_i; y_i) \\ &= \min_{\mathcal{Q} \in Q} \min_{\beta} \sum_{i \in \mathcal{Q}} d(\mu_i; y_i). \end{aligned}$$

  - All possible $\binom{n}{q}$ partitions of the data have to be fitted by the MLE (or MDE), and the MTDE is given by the partition with the minimum of the deviance residual.
- The related breakdown properties for MTDE and MTLE can be directly referred to Müller and Neykov (2003) and Neykov *et al.* (2007).

Fast algorithm

- Proposed by Rousseeuw and van Driessen (1999,2006)
- Müller and Neykov (2003) develop FAST-TLE to get an approximative MTLE.
- The fast algorithm selects a small subset of the data at first, and then the subset is augmented and updated in such a way that outliers are unlikely to be included.
- The basic idea behind the fast algorithm is to carry out a two-step iterative procedures repeatedly:
  - a trial step
  - a refinement step (the so-called concentration step)

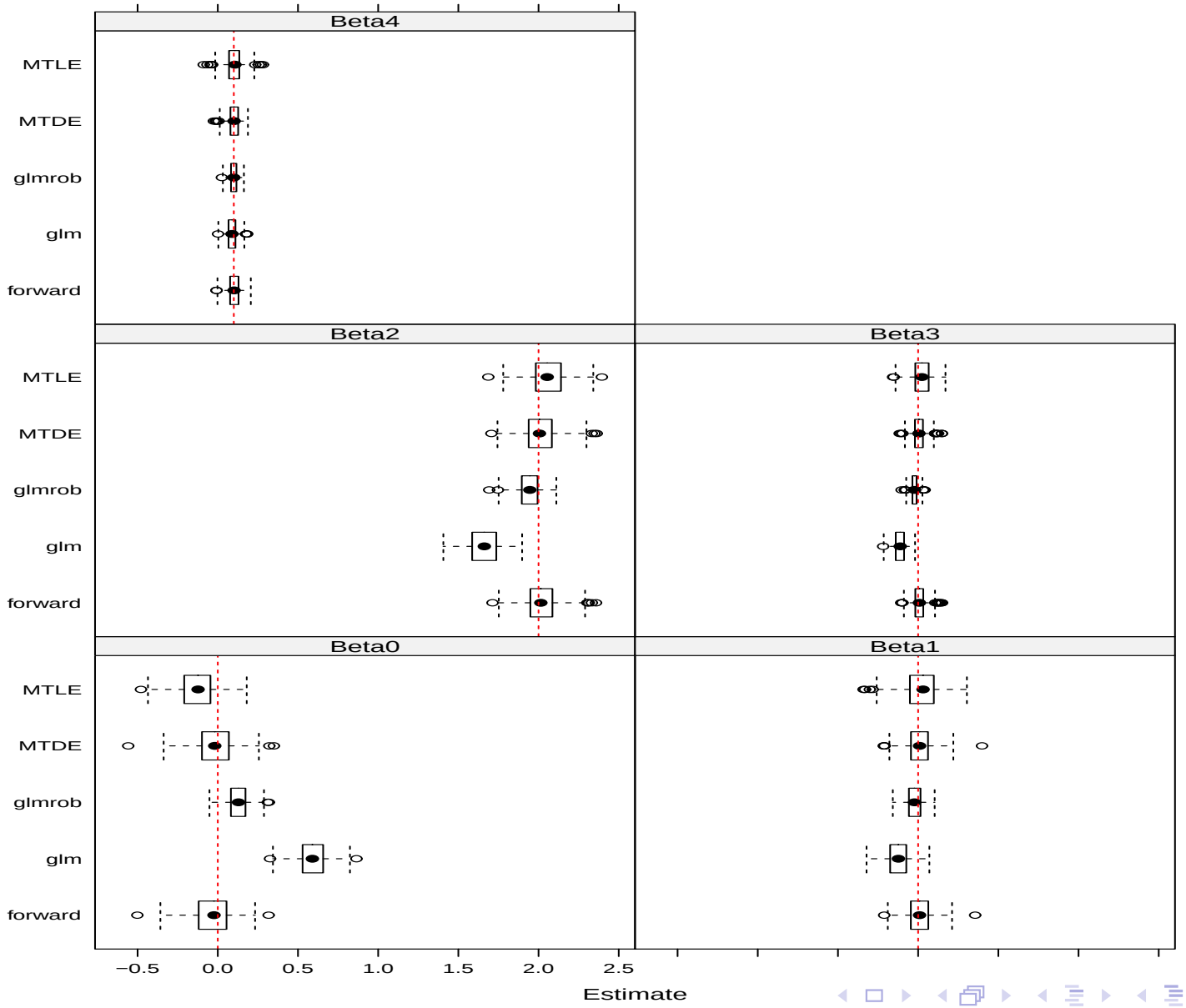
Fast algorithm for MTLE (MTDE)

- To carry out the fast algorithm for MTLE (MTDE), in the trial step a subsample of size s is selected randomly from the data and then the model is fitted to that subsample to get a trial ML estimate.
- The refinement step is designed by the so-called concentration procedure as follows:
  - (a) the observations with the q largest log-likelihood (q smallest deviance residuals) based on the current estimate are found, starting with the trial MLE as initial estimator;
  - (b) fitting the model to these q observations yields an improved fit.
  - Repeating (a) and (b) leads to an iterative procedure.
- The convergence is always guaranteed after a finite number of steps since there are only a finite many q-subsets out of $\binom{n}{q}$.
- The one with the largest value of the sum of q largest log-likelihood is then an approximate to the solution of MTLE.
- (The one with the smallest value of the sum of q smallest deviance residuals is then an approximate to the solution of MTD.)
- The resulting estimates are denoted by $\hat{\boldsymbol{\beta}}_q$ and $\hat{\alpha}_q$.

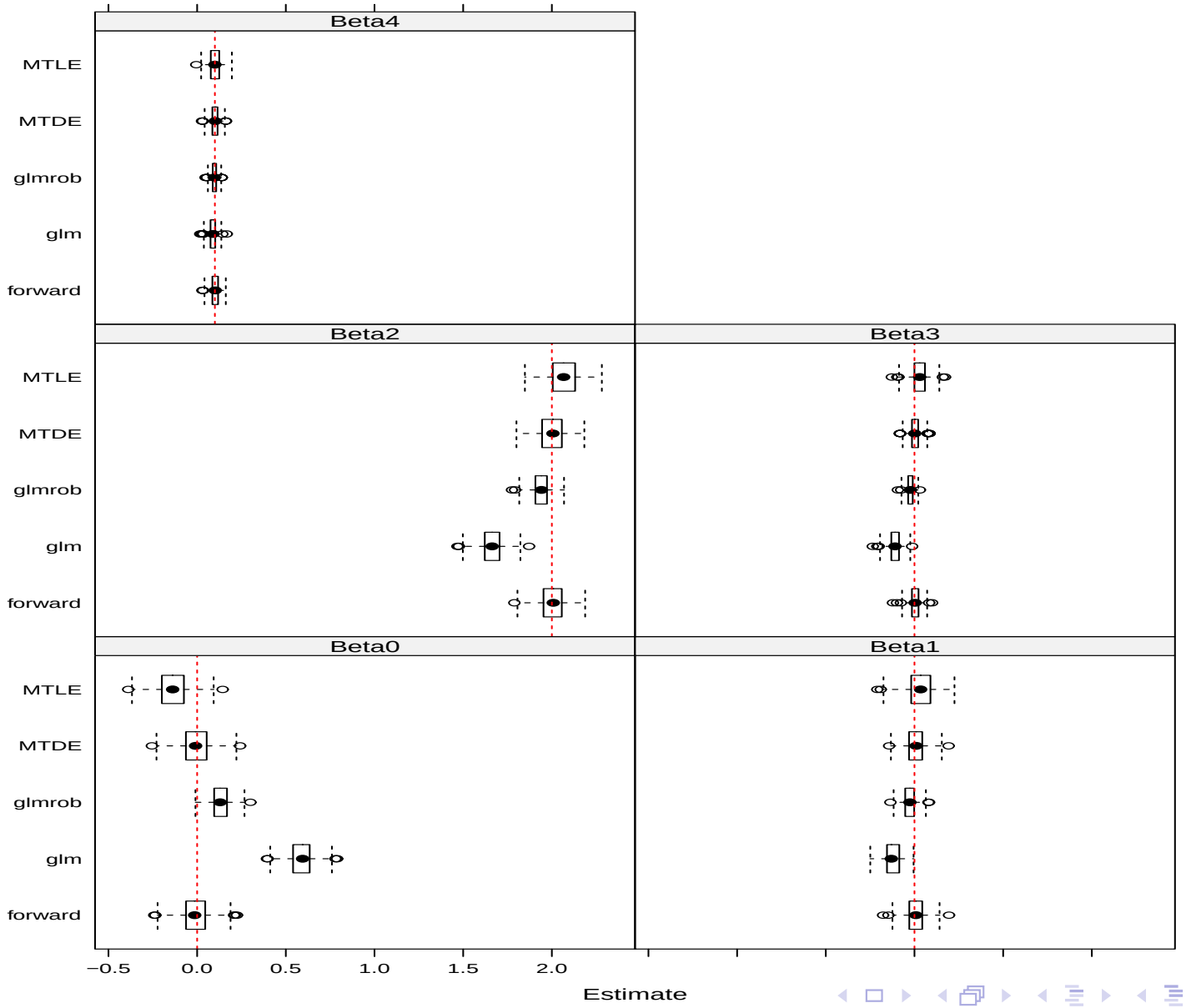
Simulation study

- The covariates are generated as follows (Hall and Shen 2009)
  - X_1 and X_2 are generated from a binomial distribution with the success rate being 0.5
  - X_3 and X_4 are generated from $N(0, 1)$
- Outlier in Y
  - Let $\mu_i = X_{1i} + 2X_{2i} + X_{3i} + 0.1X_{4i}$, $i = 1, 2, \cdots, n$.
  - The response value y_i is generated from a Poisson distribution with mean $\exp(\mu_i)$
  - The first 10% of the data are shifted to be outliers by adding 15 to these values of y_i's.
- Outlier in X
  - Let $\mu_i = 0.0X_{1i} + X_{2i} + X_{3i} + 0.1X_{4i}$, $i = 1, 2, \cdots, n$.
  - The response value y_i is generated from a Poisson distribution with mean $\exp(\mu_i)$
  - The first 10% of the data for X_3 are replaced by $X_3 + 3$.
- The sample sizes (n) 200 and 400 are considered.
- 300 replicates are used to examine the performance.

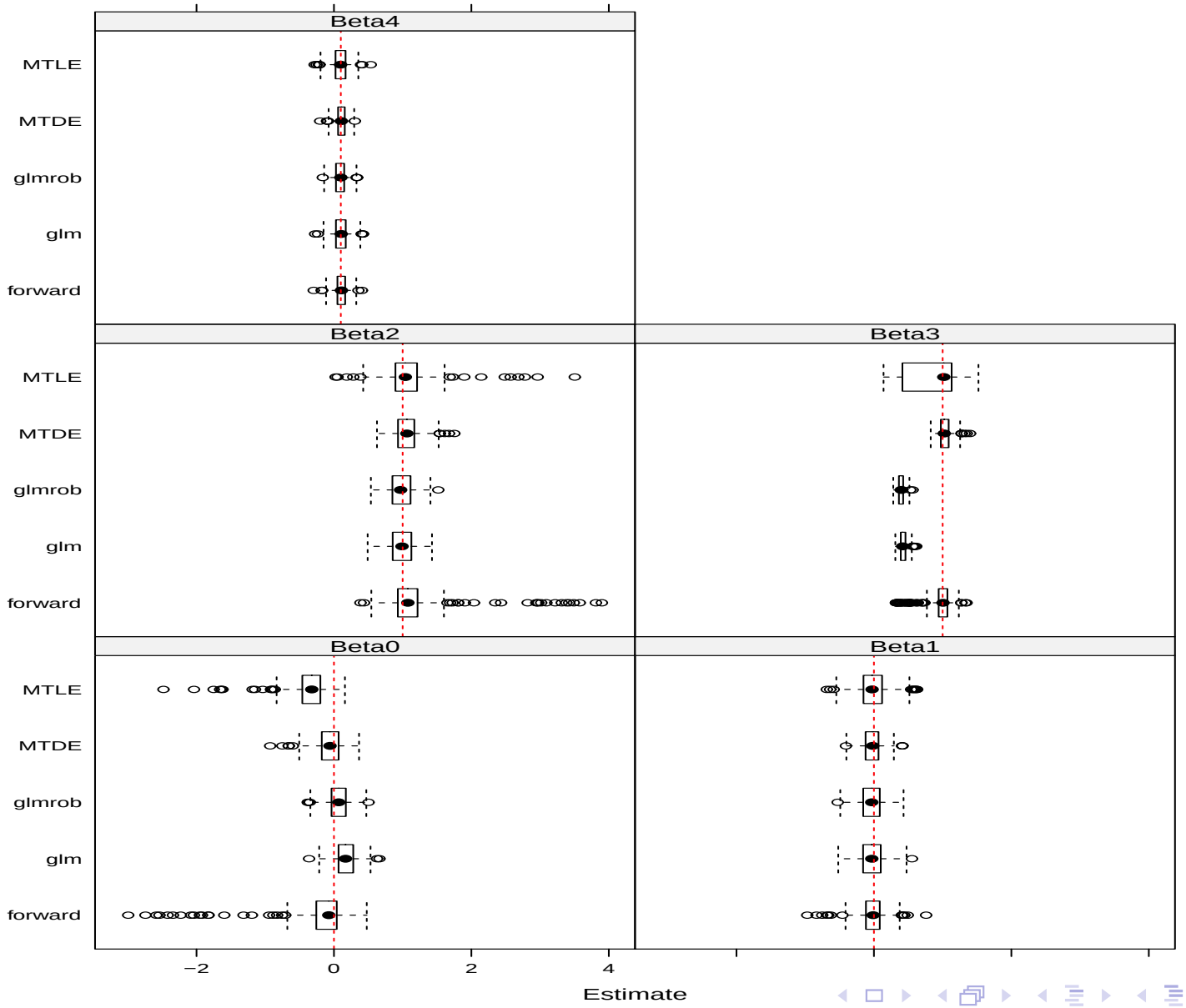
Outliers in y, $n = 200$



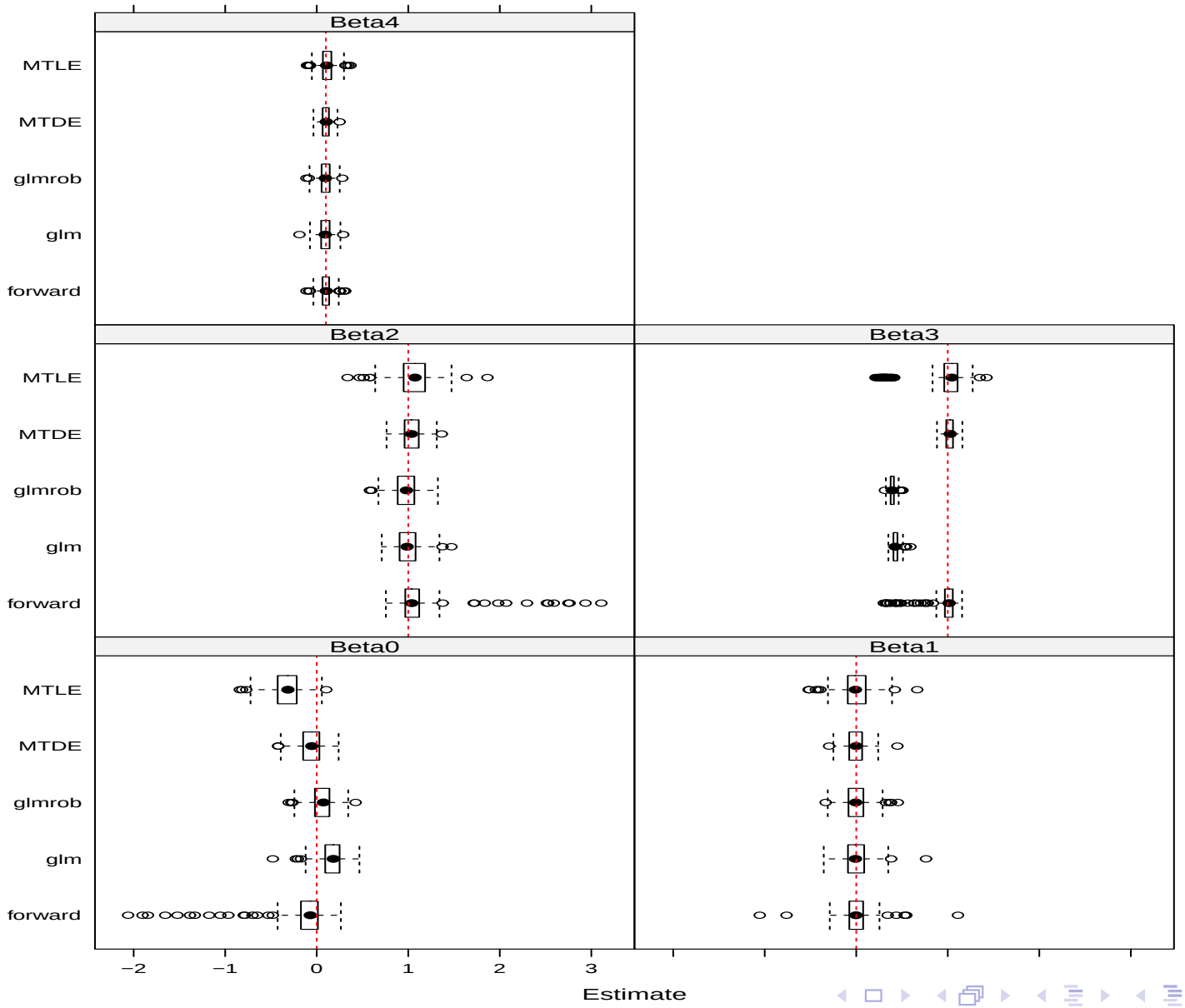
Outliers in y, $n = 400$



Outliers in X, $n = 200$

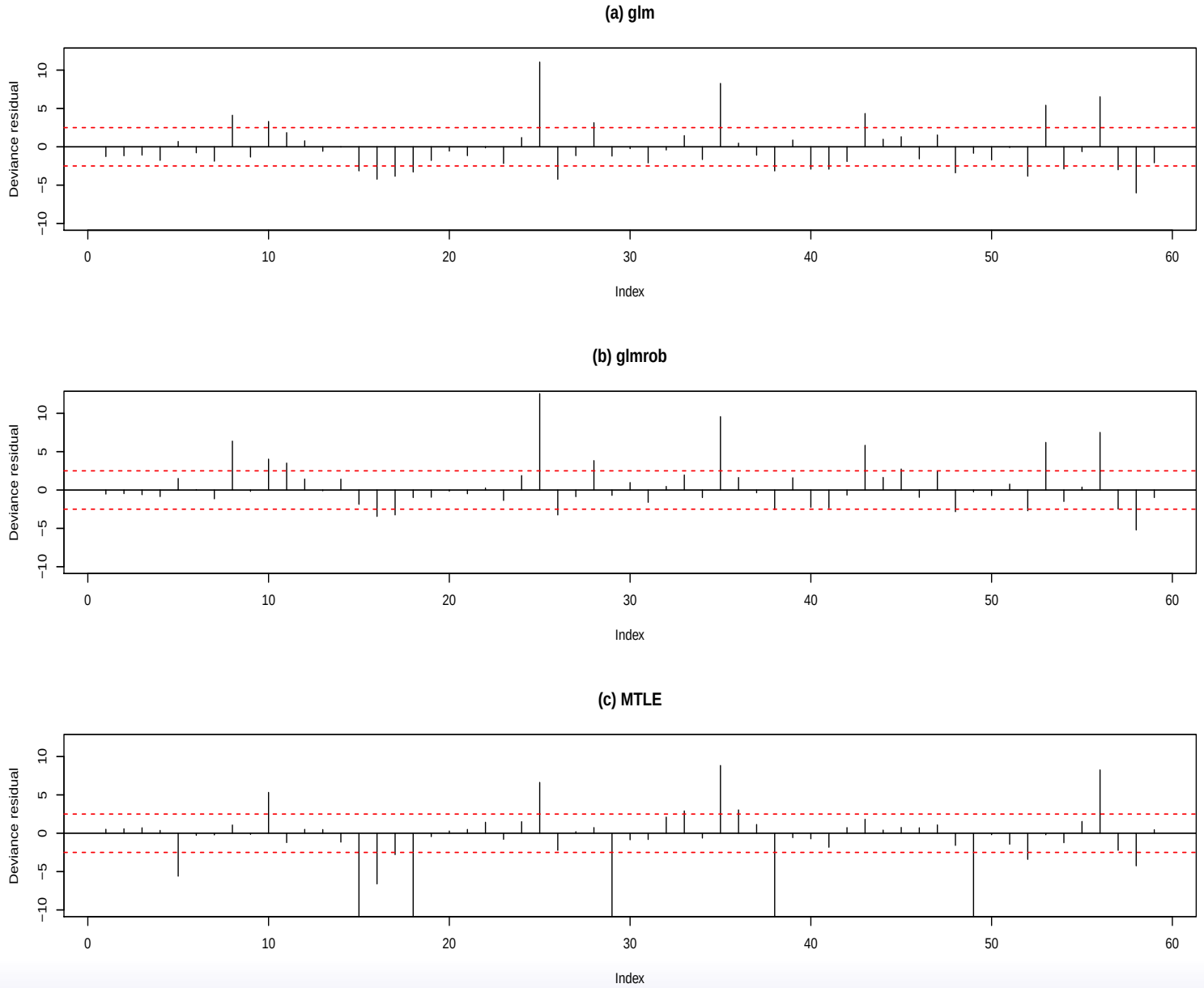


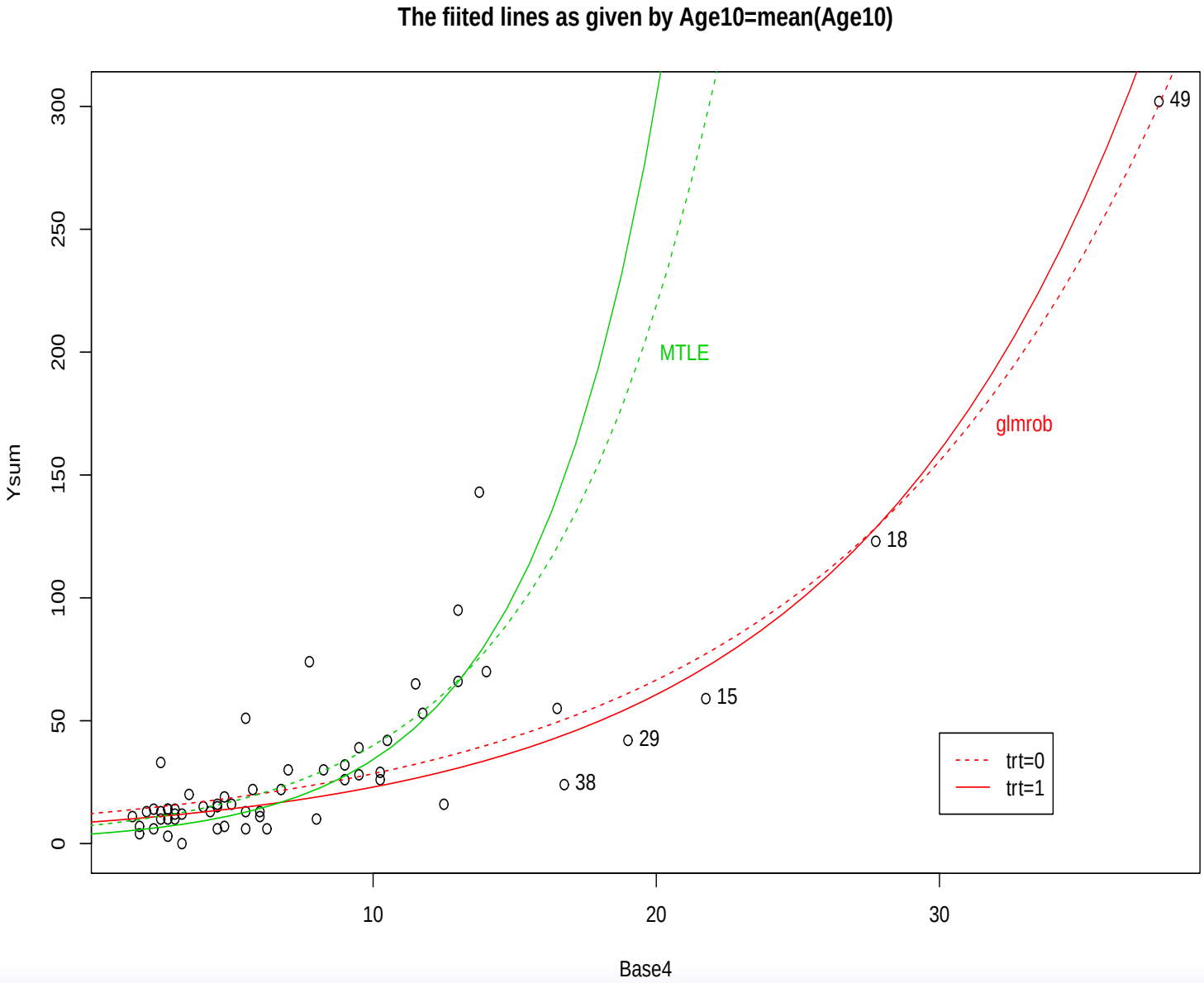
Outliers in X, $n = 400$



Epilepsy data

- Breslow (1996) used a GLM with Poisson response and log link to study the effect of drugs in epilepsy patients.
- A clinical trial of 59 patients with epilepsy, 31 of whom were randomized to receive the anti-epilepsy drug Progabide and 28 of whom received a placebo.
- The response variable is SumY: the number of attacks during four weeks in a given time interval.
- The explanatory variables are
  - Age10: patient age divided by ten
  - Base4: number of attacks in the four weeks prior to the study
  - Trt: a dummy variable that takes the values 1 or 0 if the patient received the drug or a placebo, respectively
  - Base4*Trt: to take into account the interaction between these two variables
- glmrob(Ysum $\sim$ Age10 + Base4*Trt, family=poisson, data=epilepsy) (e.g. Valodra and Yohai 2014)


(a) glm
Deviance residual
10
5
0
−5
−10
0
10
20
30
40
50
60
Index
(b) glmrob
Deviance residual
10
5
0
−5
−10
0
10
20
30
40
50
60
Index
(c) MTLE
Deviance residual
10
5
0
−5
−10
0
10
20
30
40
50
60
Index


The fiited lines as given by Age10=mean(Age10)
Ysum
Base4
0
50
100
150
200
250
300
10
20
30
MTLE
glmrob
49
18
15
29
38
trt=0
trt=1

MTLE (or MTD) deviance residual for NB

- Once the MTLE (MTD) estimates of $\boldsymbol{\beta}$ is obtained, the estimated value of the model is $\log(\hat{\mu}_{i,q}) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_q$.
- The signed square-root deviance residual for the ith observation is then

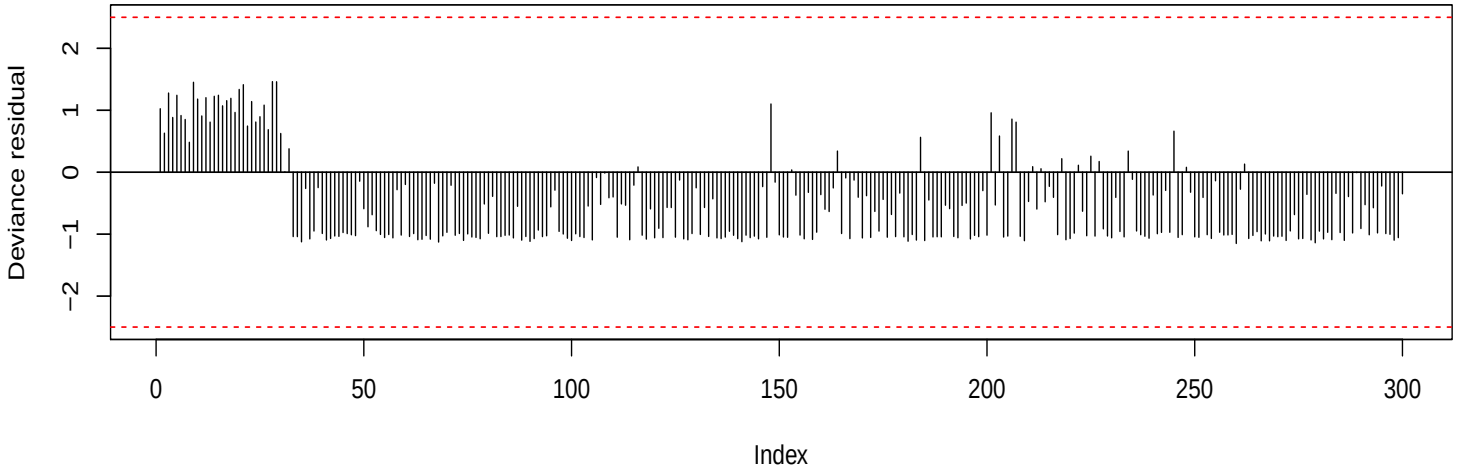
$$\begin{aligned} d_{i,q} &= d(y_i; \hat{\mu}_{i,q}, \hat{\alpha}_q) = \text{sign}(y_i - \hat{\mu}_{i,q})\sqrt{2} \\ &\quad \times \left\{ \frac{1}{\hat{\alpha}_q} \log \frac{1 + \hat{\alpha}_q \hat{\mu}_{i,q}}{1 + \hat{\alpha}_q y_i} \right. \\ &\quad \left. + y_i \log \frac{y_i(1 + \hat{\alpha}_q \hat{\mu}_{i,q})}{\hat{\mu}_{i,q}(1 + \hat{\alpha}_q y_i)} \right\}^{1/2} \end{aligned}$$

Simulated data

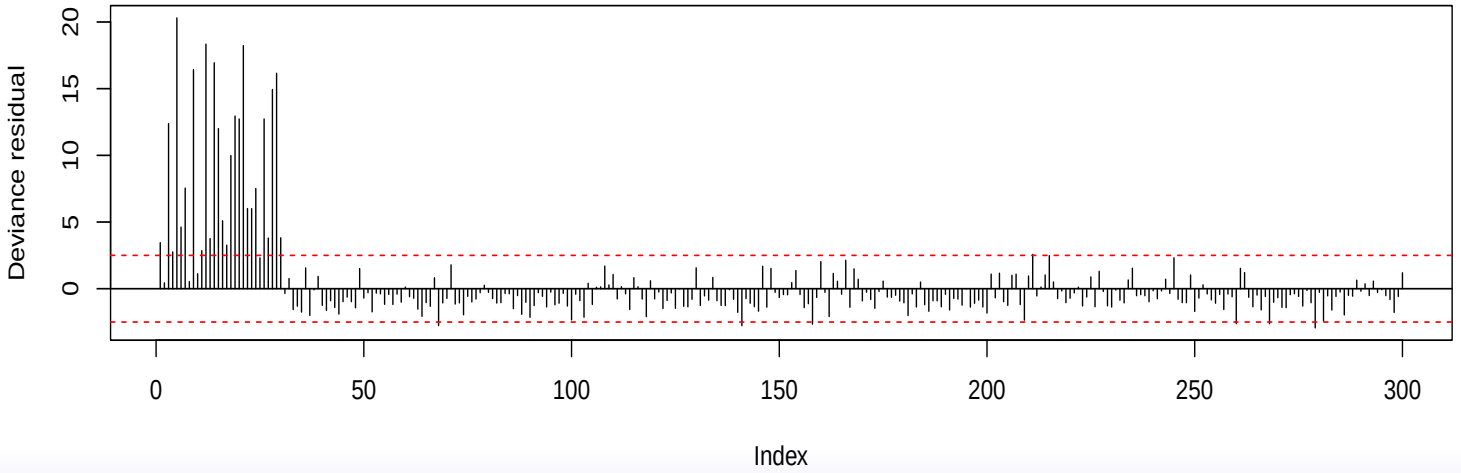
- The simulated data are generated as follows:
  - X_1, X_2, and X_3 are generated from $N(0, 1)$.
  - Let $\mu_i = x_{1i} + x_{2i} + x_{3i}$, $i = 1, 2, \cdots, 300$.
  - The response value y_i is generated from a negative binomial distribution with mean $\exp(\mu_i)$ and dispersion parameter $\alpha = 3.33$.
  - The first 30 observations (10% of the data) are shifted to be outliers by adding 50 to these values of y_i's ($i = 1, 2, \cdots, 30$).

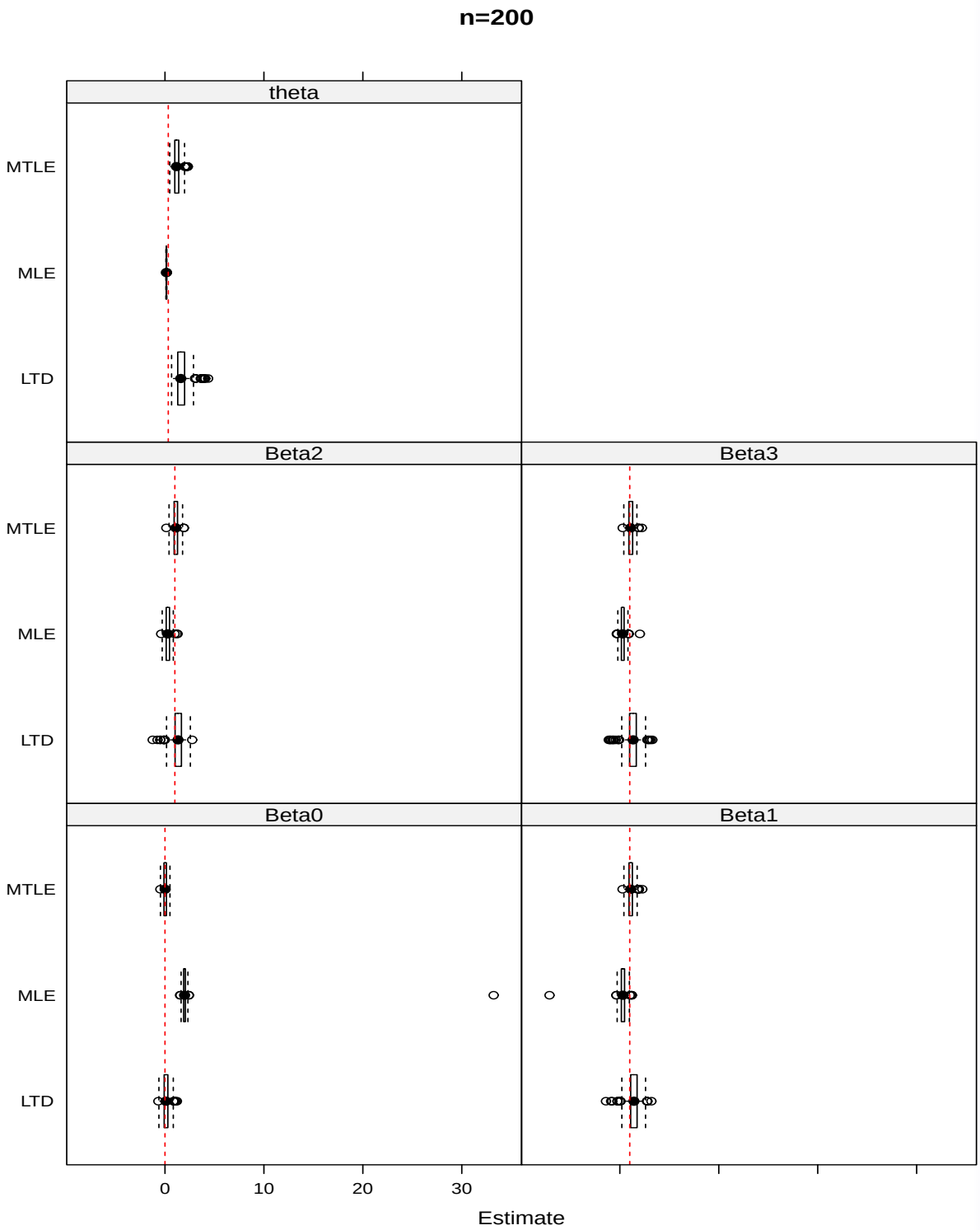
Simulated data: Deviance residual plot

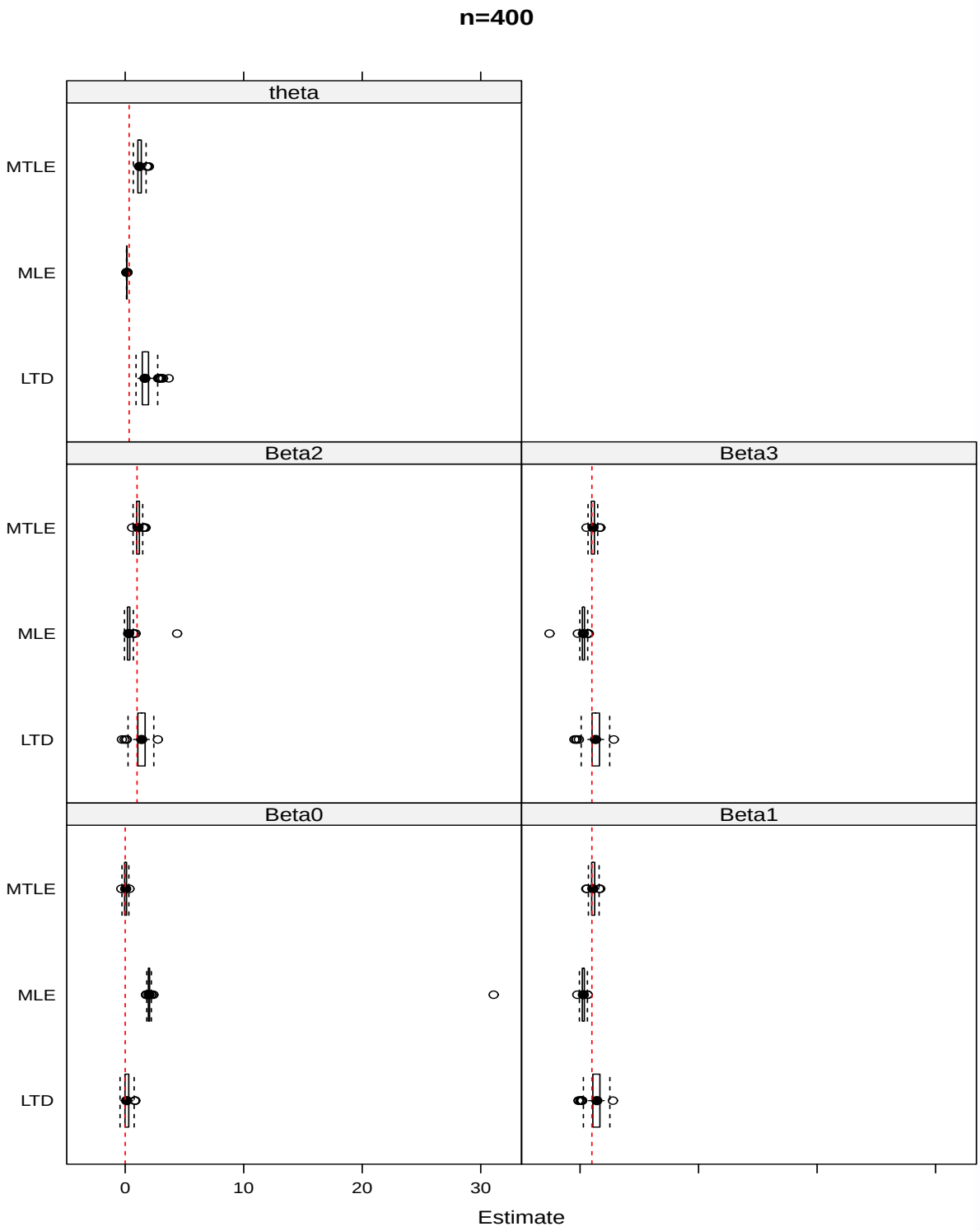
(a) MLE



(b) MTLE




n=200
theta
Beta2
Beta3
Beta0
Beta1
MTLE
MLE
LTD
0
10
20
30
Estimate


n=400
theta
Beta2
Beta3
Beta0
Beta1
MTLE
MLE
LTD
0
10
20
30
Estimate

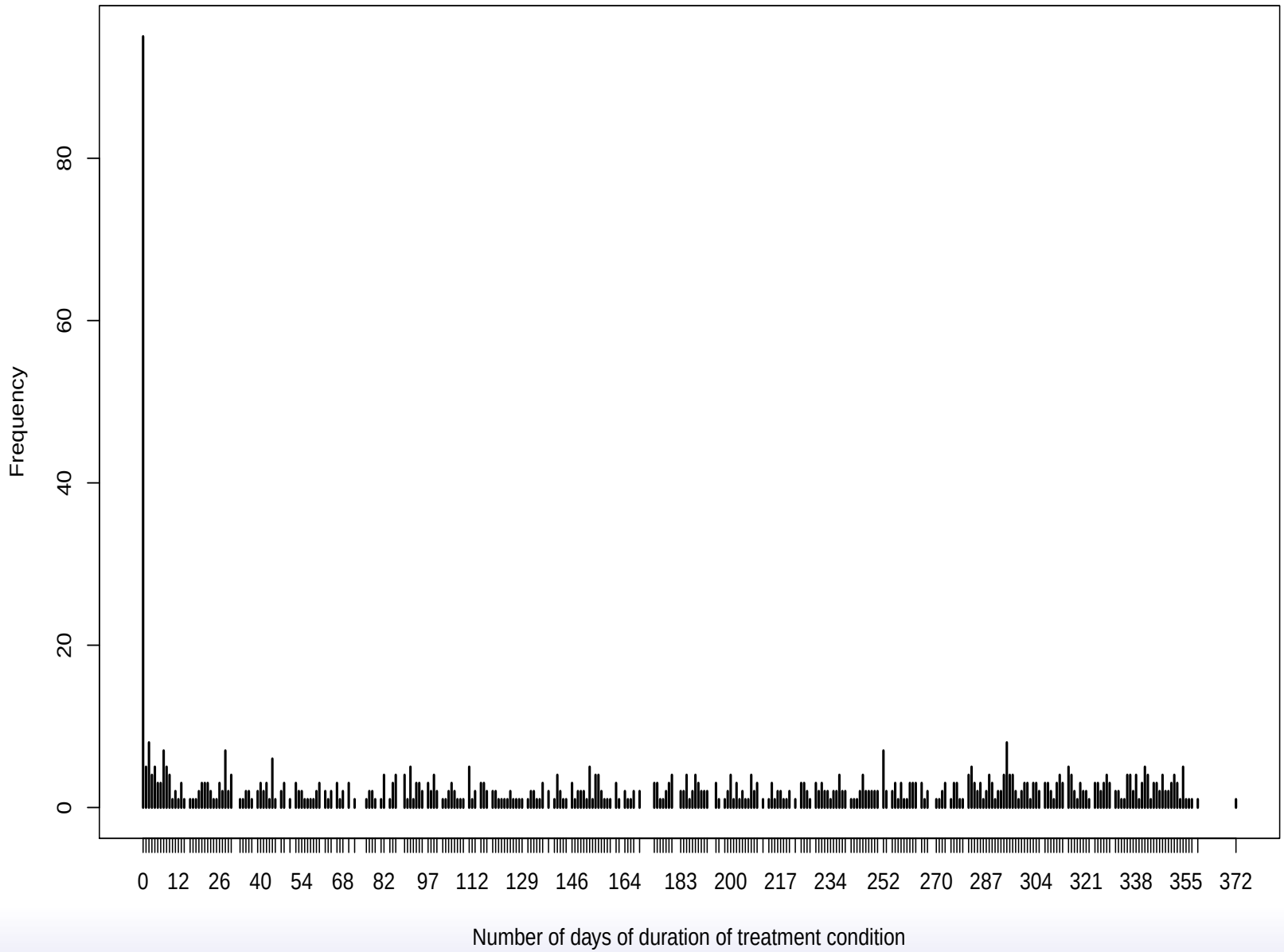
Ischemic heart disease

- Theses data were collected by a health insurance plan and provide information concerning 788 subscribers who had made claims resulting from ischemic (coronary) heart disease.
- Kutner, Nachtsheim, and Neter (2008, 4th ed., pages 683-4.)

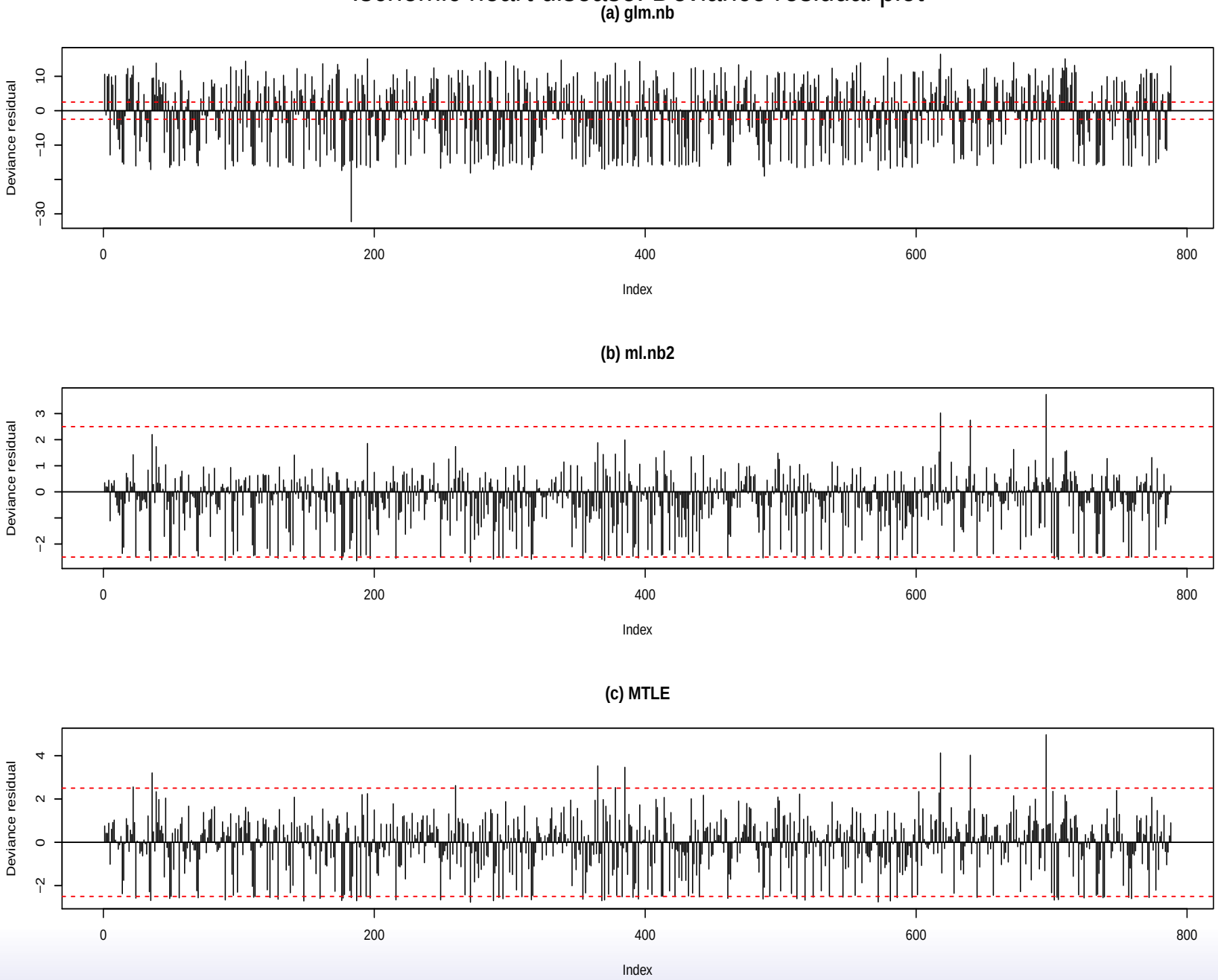
Variable	Description
Total cost	Total cost claims by subscriber (dollars)
Age	Age of subscriber (years)
Gender	Gender of subscriber: 1 if male; 0 female
Intervention	Total number of interventions or procedures carried out
Drug	Number of tracked drugs prescribed
Emergency	Number of emergency room visits
Complication	Number of other complications arose during heart disease treatment
Comorbidity	Number of other diseases that the subscriber had during period
Duration	Number of days of duration of treatment condition

- We here consider the following model

$$\begin{aligned} \textit{Duration} \quad \sim \quad & \textit{Age} + \textit{Gender} + \textit{Intervention} + \textit{Drug} + \textit{Emergency} + \\ & \textit{Complication} + \textit{Comorbidity} \end{aligned}$$


Frequency
0
20
40
60
80
0 12 26 40 54 68 82 97 112 129 146 164 183 200 217 234 252 270 287 304 321 338 355 372
Number of days of duration of treatment condition

Ischemic heart disease: Deviance residual plot



Ischemic heart disease: estimation results

	glm.nb			ml.nb2			MTLE (75%)		
Variable	β	StdErr	t	β	StdErr	t	β	StdErr	t
Intercept	4.095	0.026	155.84	0.151	0.336	0.45	-0.324	0.431	-0.75
Age	0.011	0.000	26.08	0.073	0.006	12.42	0.072	0.007	9.90
Gender	0.053	0.007	7.94	0.262	0.118	2.22	-0.031	0.122	-0.25
Intervention	0.016	0.000	33.10	0.035	0.010	3.38	0.037	0.010	3.75
Drug	0.002	0.003	0.55	0.230	0.068	3.36	0.098	0.053	1.84
Emergency	0.017	0.001	12.91	0.003	0.023	0.15	0.059	0.023	2.59
Complication	0.081	0.010	7.77	0.888	0.317	2.80	-0.273	0.251	-1.09
Comorbiditiy	0.037	0.000	118.84	0.073	0.010	7.10	0.109	0.012	9.38
α	0.000			1.728	0.092	18.77	1.421	0.086	16.57