A New Method for Estimating Overdispersion in Count Data

David Fletcher

May 5, 2011

Overview

- Motivation
- Example
- Overdispersion in count data
- Current methods for estimating overdispersion
- Quadratic estimating equations
- New method
- Simulations
- Further work

Motivation

- Mark-recapture data (multinomial likelihood)
- Lack-of-fit diagnostics: how can we improve the model?
- Overdispersion: variation greater than predicted

- Count data: estimating overdispersion tricky when mean is low
- Bootstrap-based method?
- Farrington's modification to Pearson's lack-of-fit statistic
- Estimating overdispersion vs testing lack-of-fit

Example: sealion bycatch

	Sealions killed			
	0	1	2	3
Spring	20	1	0	0
Summer	87	3	0	0
Autumn	125	27	4	2

- Bycatch of New Zealand sealions in squid fishery (1990 – 1996)
- Sample unit: all tows by one vessel in six hours
- Offset can allow for number of tows per sample unit (1 – 40)
- Low mean count

Overdispersion in count data

- Poisson model

$$Y_i \sim Poisson\left(\mu_i\right)$$
$$E\left(Y_i\right) = V\left(Y_i\right) = \mu_i$$
$$g\left(\mu_i\right) = X\beta$$

  g is the link function, X is the $n \times p$ model matrix and β is the vector of model parameters

- Variance often greater than the mean (overdispersion)
- SEs too small and overly-complex models selected
- Use estimate of ϕ to adjust SEs and model selection

Overdispersion in count data

- Quasi-Poisson approach (special case of quasi-liklihood):

$$E(Y_i) = \mu_i \qquad V(Y_i) = \phi\mu_i$$

  where ϕ (> 1) is the dispersion parameter
- Leads to same $\hat{\beta}$ as Poisson model

- Negative binomial model:

$$E(Y_i) = \mu_i \qquad V(Y_i) = \mu_i + \frac{\mu_i^2}{k_i}$$

- $k_i = \mu_i / (\phi - 1) \Rightarrow V(Y_i) = \phi\mu_i$

Current methods for estimating ϕ

- Quasi-Poisson approach:

$$E(Y_i) = \mu_i \qquad V(Y_i) = \phi\mu_i$$

- Wedderburn 1974 proposed the current standard:

$$\hat{\phi}_X = \frac{X^2}{n-p}$$

where

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

- Prone to problems when μ_i small: highly variable

Current methods for estimating ϕ

- Farrington (1995) proposed:

$$\hat{\phi}_F = \frac{X^2 - \sum_i \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}}{n-p} = \frac{\sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} - \sum_i \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}}{n-p}$$

- Asymptotic theory suggests $\hat{\phi}_F$ better than $\hat{\phi}_X$, especially when μ_i small: less variable and lower correlation with $\hat{\beta}$

Quadratic estimating equations

- Crowder (1987) proposed the general use of quadratic estimating equations. For overdispersed count data these are:

$$g_r(\theta) = \sum_i \left[a_i^{(r)} (y_i - \mu_i) + b_i^{(r)} \left\{ (y_i - \mu_i)^2 - \phi \mu_i \right\} \right] = 0$$

  for $r = 1, \ldots, p+1$, where $\theta = (\beta, \phi)^T$ and the $a_i^{(r)}$, $b_i^{(r)}$ are functions of θ
- Special cases: least squares and quasi-likelihood ($b_i^{(r)} = 0$)
- Partly motivated by issues with quasi-likelihood, including desire to use information on β contained in the variance

Quadratic estimating equations

- For overdispersed count data these are:

$$g_r(\theta) = \sum_i \left[a_i^{(r)} (y_i - \mu_i) + b_i^{(r)} \left\{ (y_i - \mu_i)^2 - \phi \mu_i \right\} \right] = 0$$

  for $r = 1, \ldots, p+1$, where $\theta = (\beta, \phi)^T$ and the $a_i^{(r)}$, $b_i^{(r)}$ are functions of θ
- Optimal choices available for $a_i^{(r)}$ and $b_i^{(r)}$ IF we specify third and fourth cumulants
- Optimal in the sense that

$$V_0\left(\hat{\theta}^*\right) - V_0\left(\hat{\theta}\right)$$

  is positive-semidefinite, where $\hat{\theta}^*$ is obtained using any other choice of $a_i^{(r)}$ and $b_i^{(r)}$

Optimal $a_i^{(r)}$ and $b_i^{(r)}$

For $r = 1, \ldots, p$

$$
\begin{aligned}
a_i^{(r)} &= \left(\frac{\phi^{1/2} \mu_i^{-1/2} \gamma_1 - \gamma_2 - 2}{\phi \mu_i \left(\gamma_2 - \gamma_1^2 + 2 \right)} \right) \partial \mu_i / \partial \beta_r \\
b_i^{(r)} &= \left(\frac{\gamma_1 - \phi^{1/2} \mu_i^{-1/2}}{\phi^{3/2} \mu_i^{3/2} \left(\gamma_2 - \gamma_1^2 + 2 \right)} \right) \partial \mu_i / \partial \beta_r \\
a_i^{(p+1)} &= \frac{\gamma_1 \phi^{-3/2} \mu_i^{-1/2}}{\gamma_2 - \gamma_1^2 + 2} \\
b_i^{(p+1)} &= -\frac{\phi^{-2} \mu_i^{-1}}{\gamma_2 - \gamma_1^2 + 2}
\end{aligned}
$$

$\gamma_{1i} = \kappa_{3i} / \kappa_{2i}^{3/2}$, $\gamma_{2i} = \kappa_{4i} / \kappa_{2i}^2$ and κ_{ji} is the jth cumulant of Y_i

$$\left(\kappa_{1i} = \mu_i \qquad \kappa_{2i} = \mu_{2i} \qquad \kappa_{3i} = \mu_{3i} \qquad \kappa_{4i} = \mu_{4i} - 3\mu_{2i}^2 \right)$$

New method of estimating ϕ

- If there were a Quasi-Poisson "distribution" in the exponential family (i.e. with $V(Y_i) = \phi\mu_i$) it would satisfy

$$\kappa_{3i} = \phi^2 \kappa_{3i}^* \qquad \text{and} \qquad \kappa_{4i} = \phi^3 \kappa_{4i}^*$$

  where $\kappa_{3i}^* = \mu_i$ and $\kappa_{4i}^* = \mu_i$ are the third and fourth cumulants of the Poisson(μ_i) distribution

- If we assume these third and fourth cumulants, the optimal $a_i^{(r)}$ and $b_i^{(r)}$ lead to the same $\hat{\beta}$ as for the Quasi-Poisson (and Poisson) approach and also give

$$\tilde{\phi} = \frac{X^2}{\sum_i \frac{y_i}{\hat{\mu}_i}}$$

Comparison of methods

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

$$\hat{\phi}_X = \frac{X^2}{n-p}$$

$$\hat{\phi}_F = \frac{X^2 - \sum_i \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i}}{n-p}$$

$$\tilde{\phi} = \frac{X^2}{\sum_i \frac{y_i}{\hat{\mu}_i}}$$

Example: sealion bycatch

	Sealions killed			
	0	1	2	3
Spring	20	1	0	0
Summer	87	3	0	0
Autumn	125	27	4	2

- Poisson model with *season* as a predictor
- Log(number of tows for that sample unit) as an offset
- No evidence of zero-inflation relative to a Poisson model

$$\hat{\phi}_X = 2.17 \qquad \hat{\phi}_F = 1.19 \qquad \tilde{\phi} = 1.09$$

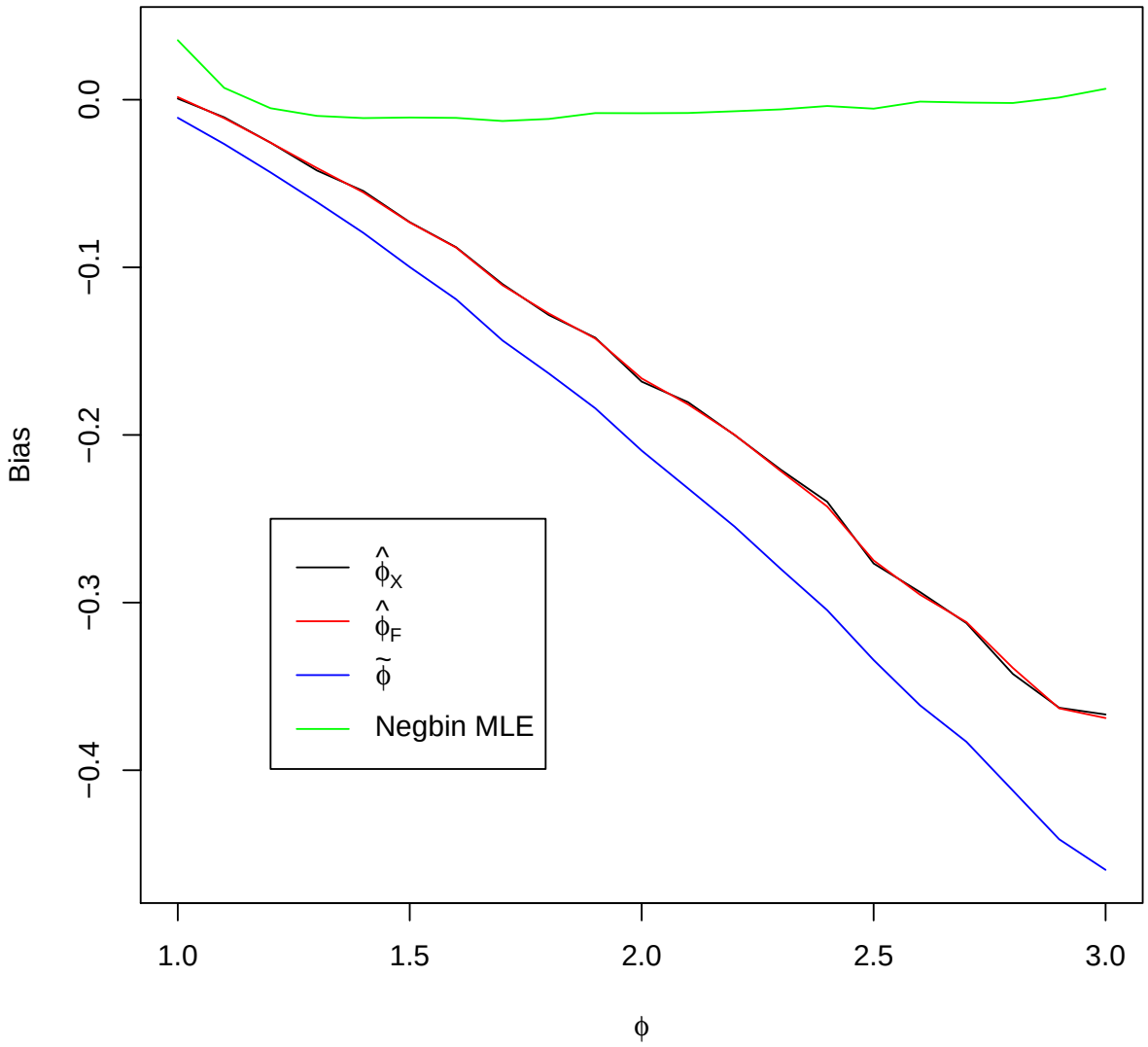
Simulations

- Based on the bycatch example (same sample sizes and number of tows per sample unit)
- β set equal to $\hat{\beta}$
- Specified true value for ϕ between 1 and 3

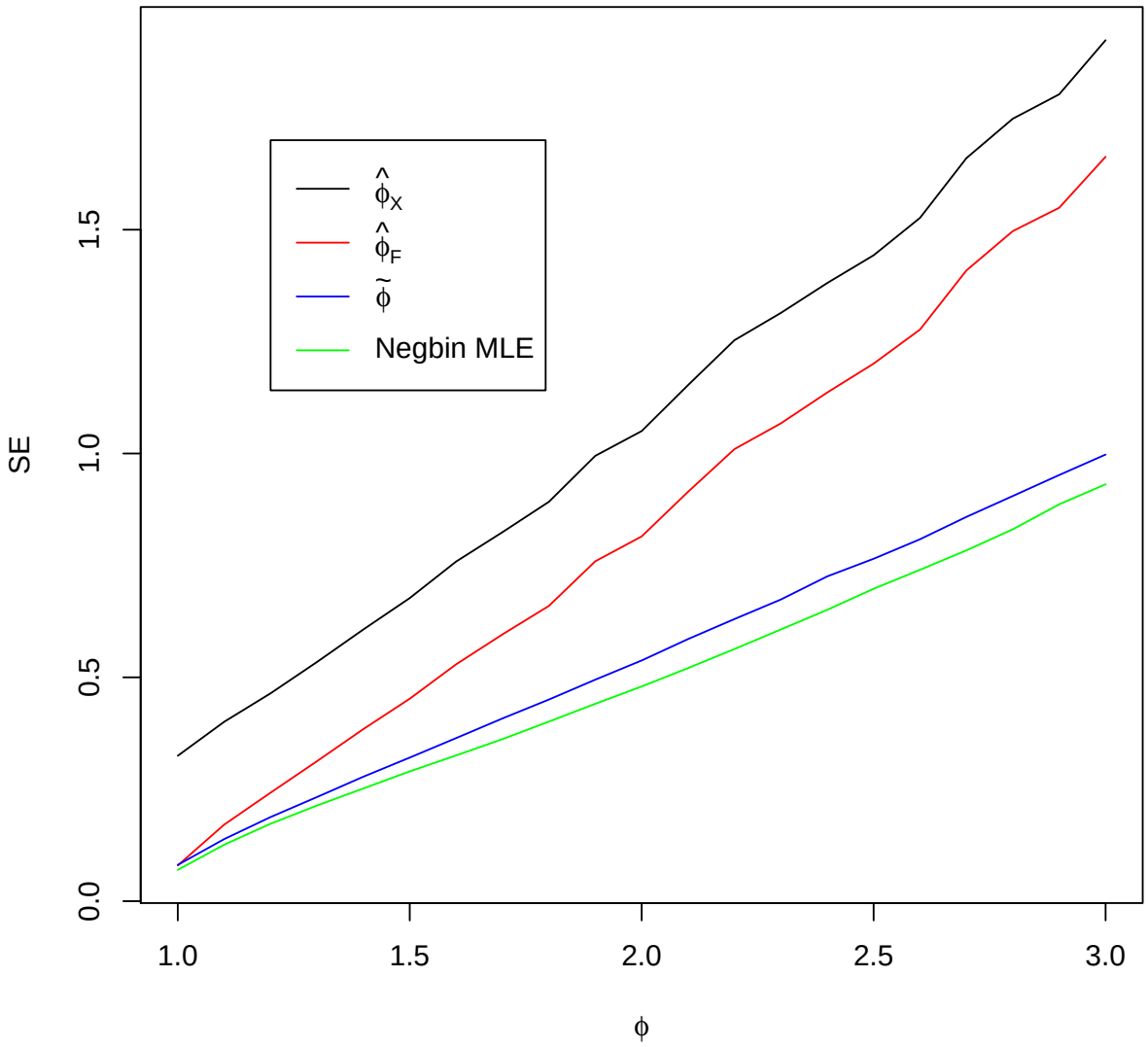
- Generated data using one of three models:

  1. Negative binomial with $V(Y_i) = \phi\mu_i$, i.e. $k_i = \mu_i / (\phi - 1)$
  2. Neyman Type A model: $Y_i = \sum_{j=1}^{N_i} Z_{ij}$, where $Z_{ij} \sim Poisson(\phi - 1)$ and $N_i \sim Poisson\left(\frac{\mu_i}{\phi - 1}\right)$
  3. For $\phi = 1$, $Y_i \sim Poisson(\mu_i)$

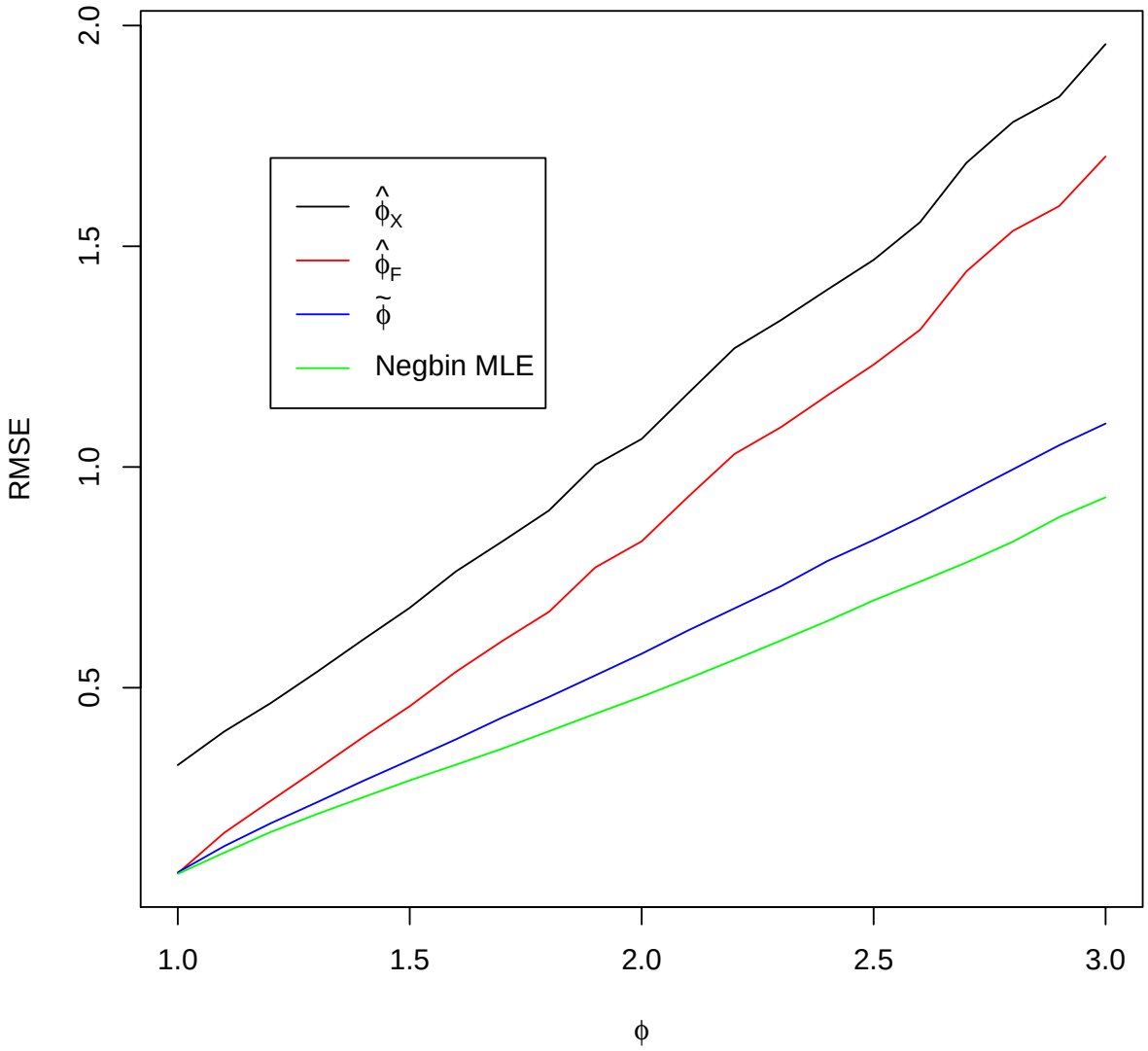
Simulation results: Negative binomial



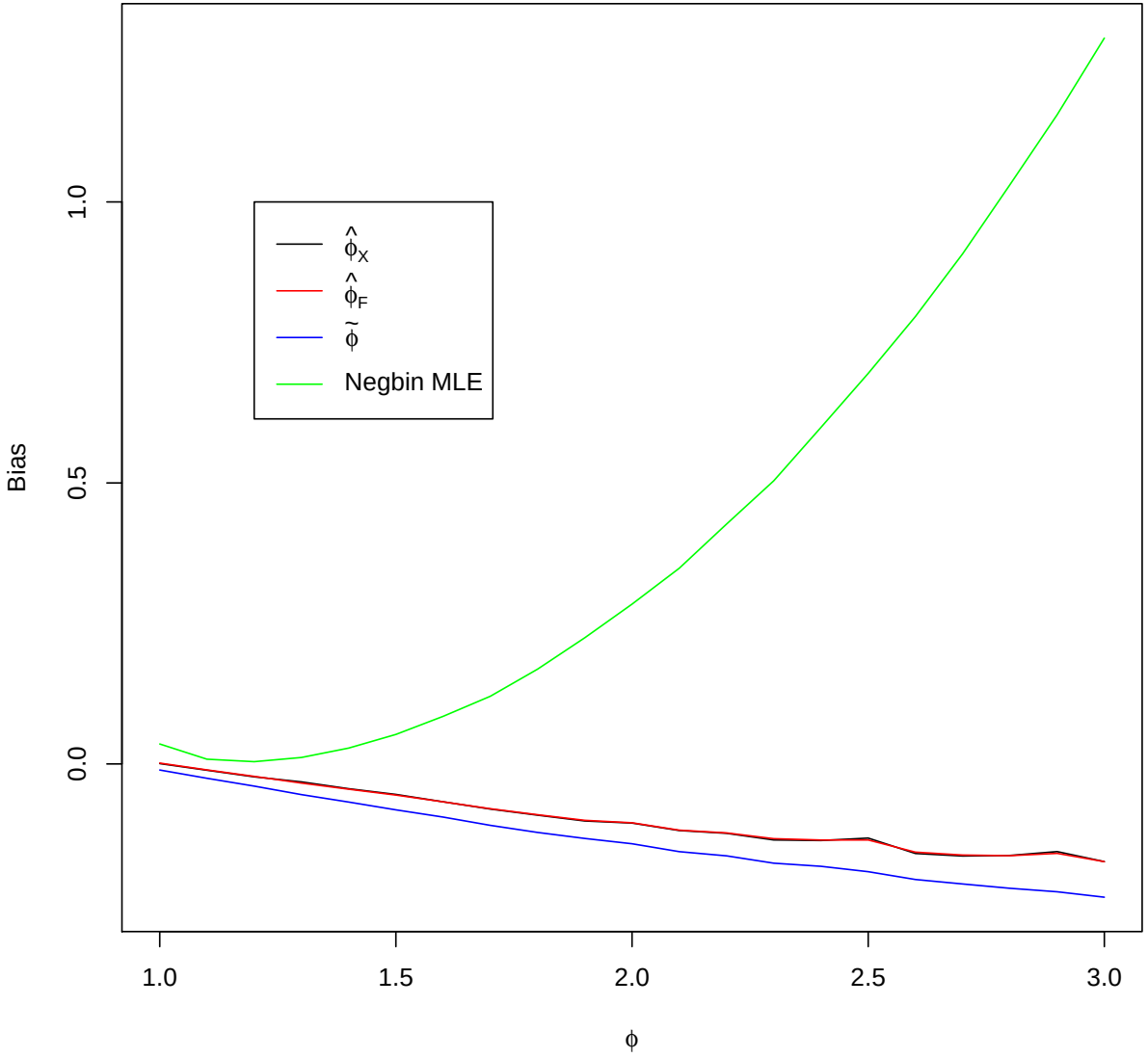
Simulation results: Negative binomial



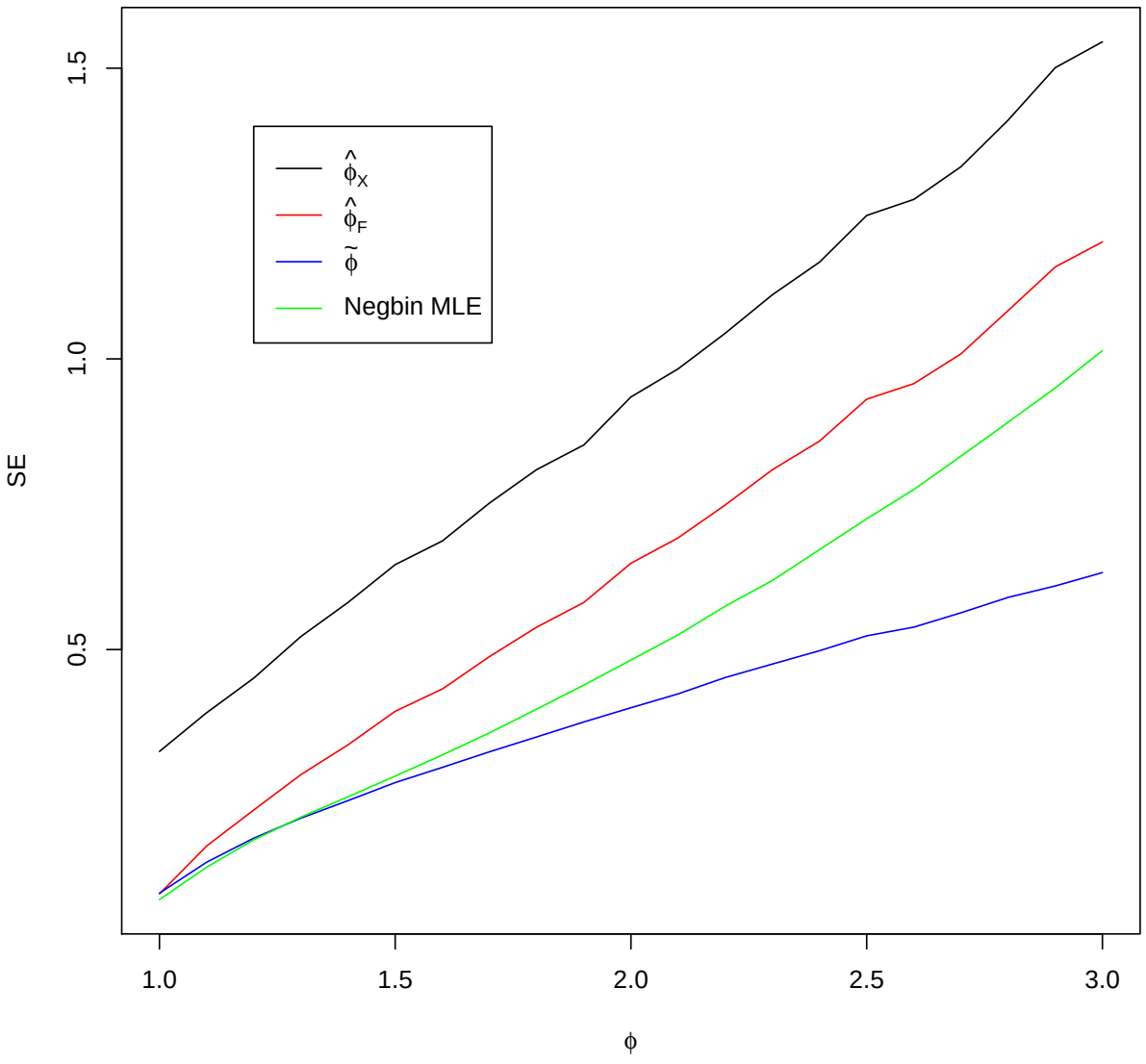
Simulation results: Negative binomial



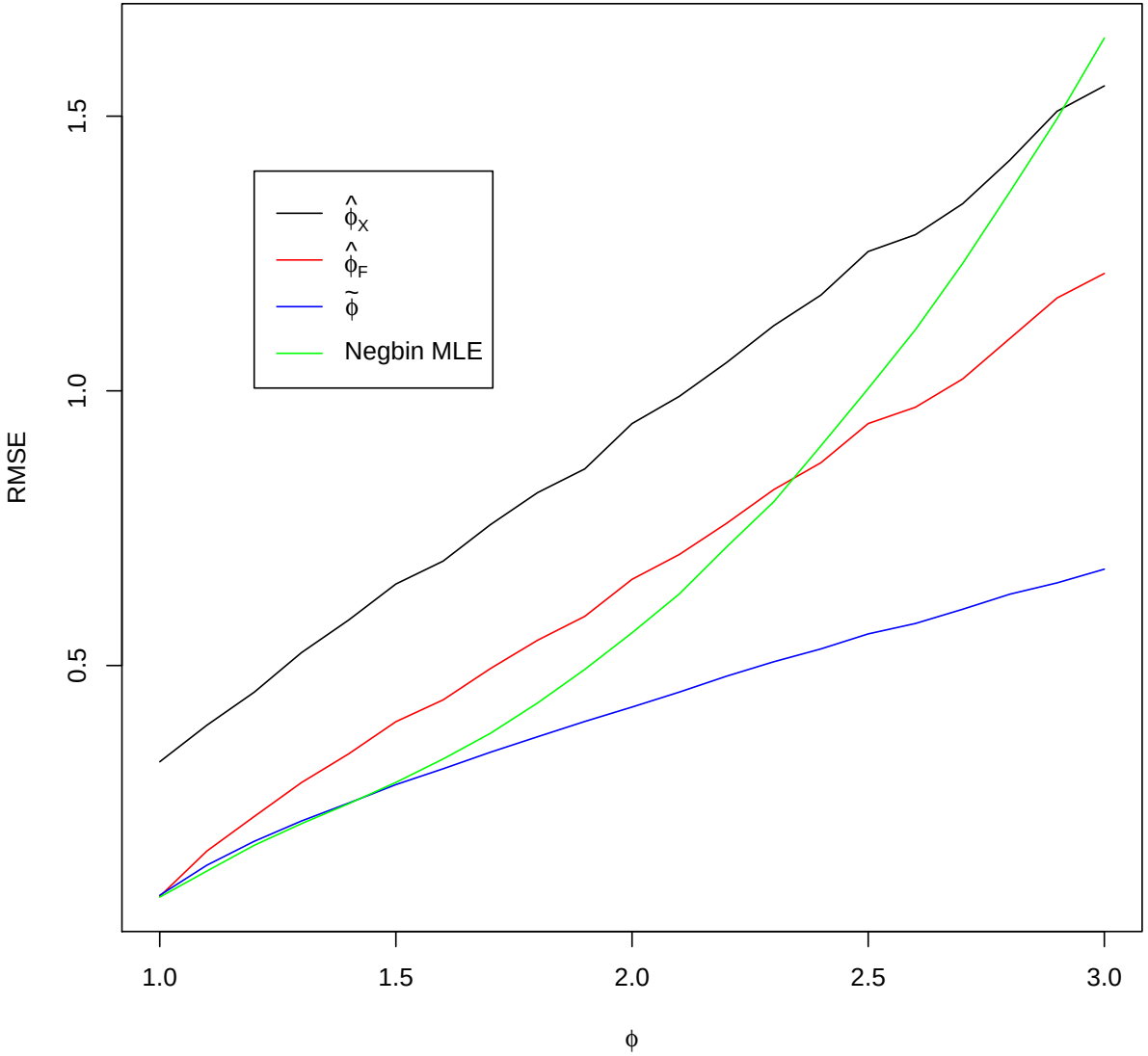
Simulation results: Neyman Type A



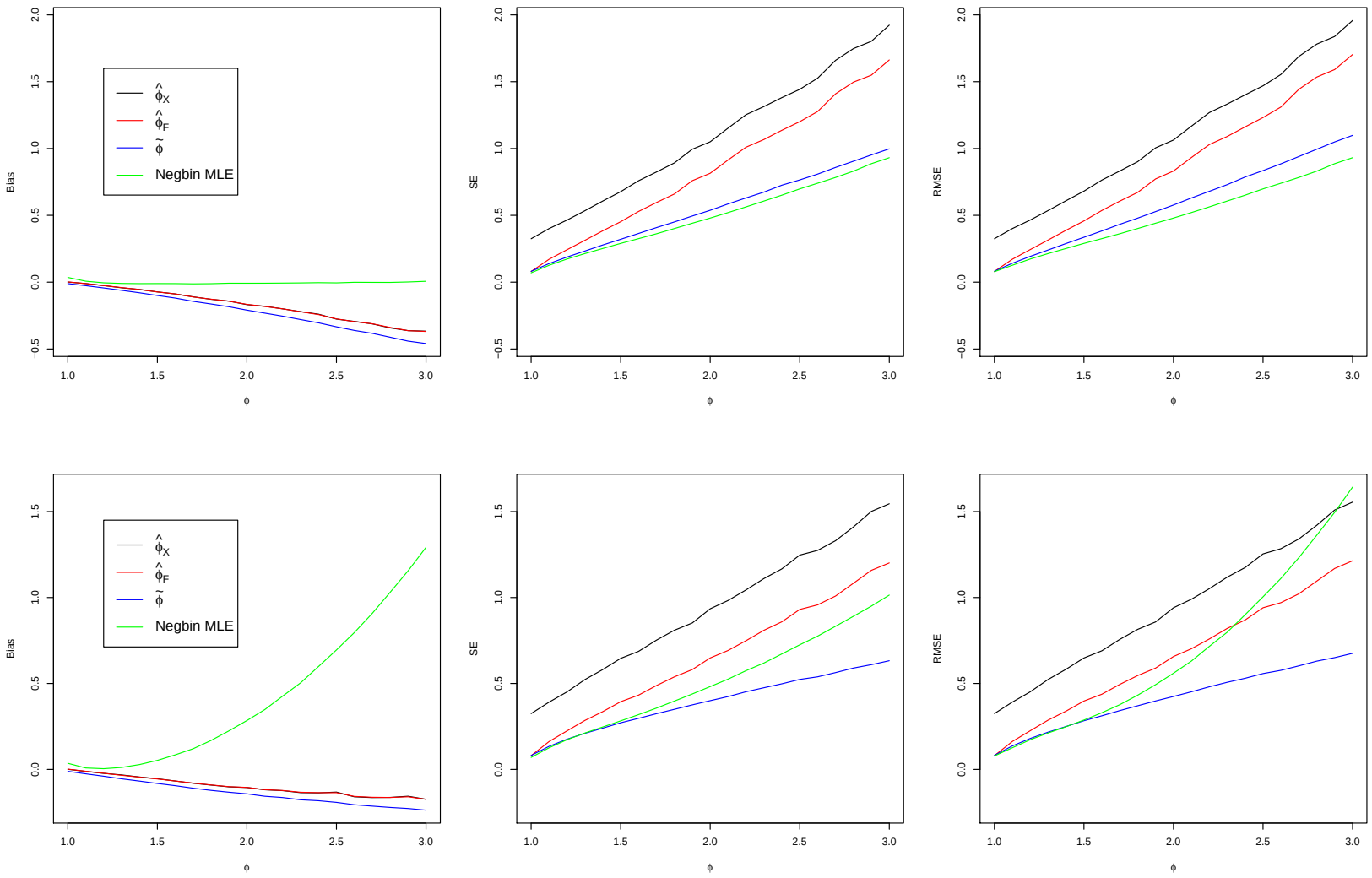
Simulation results: Neyman Type A



Simulation results: Neyman Type A



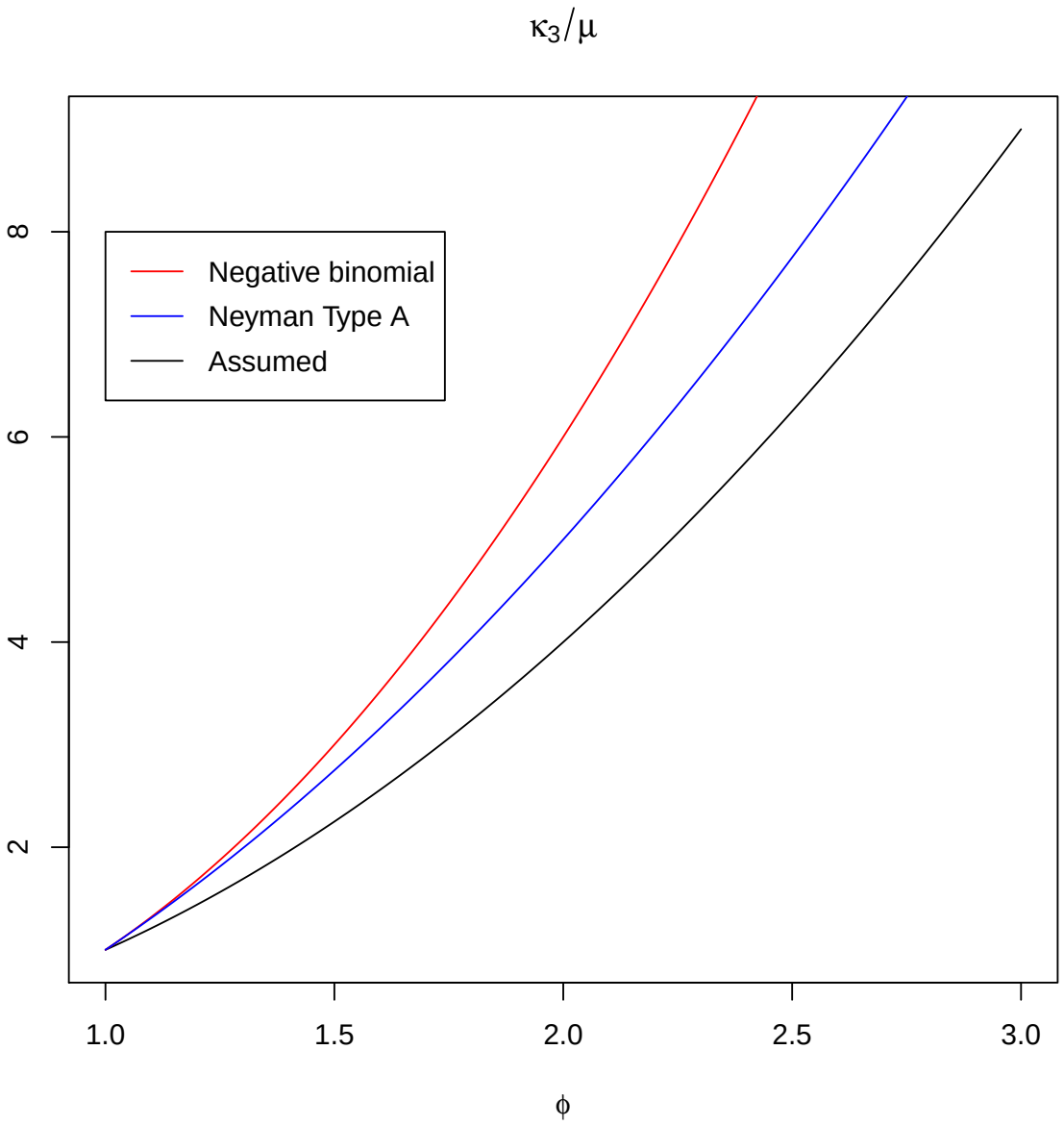
Simulation results



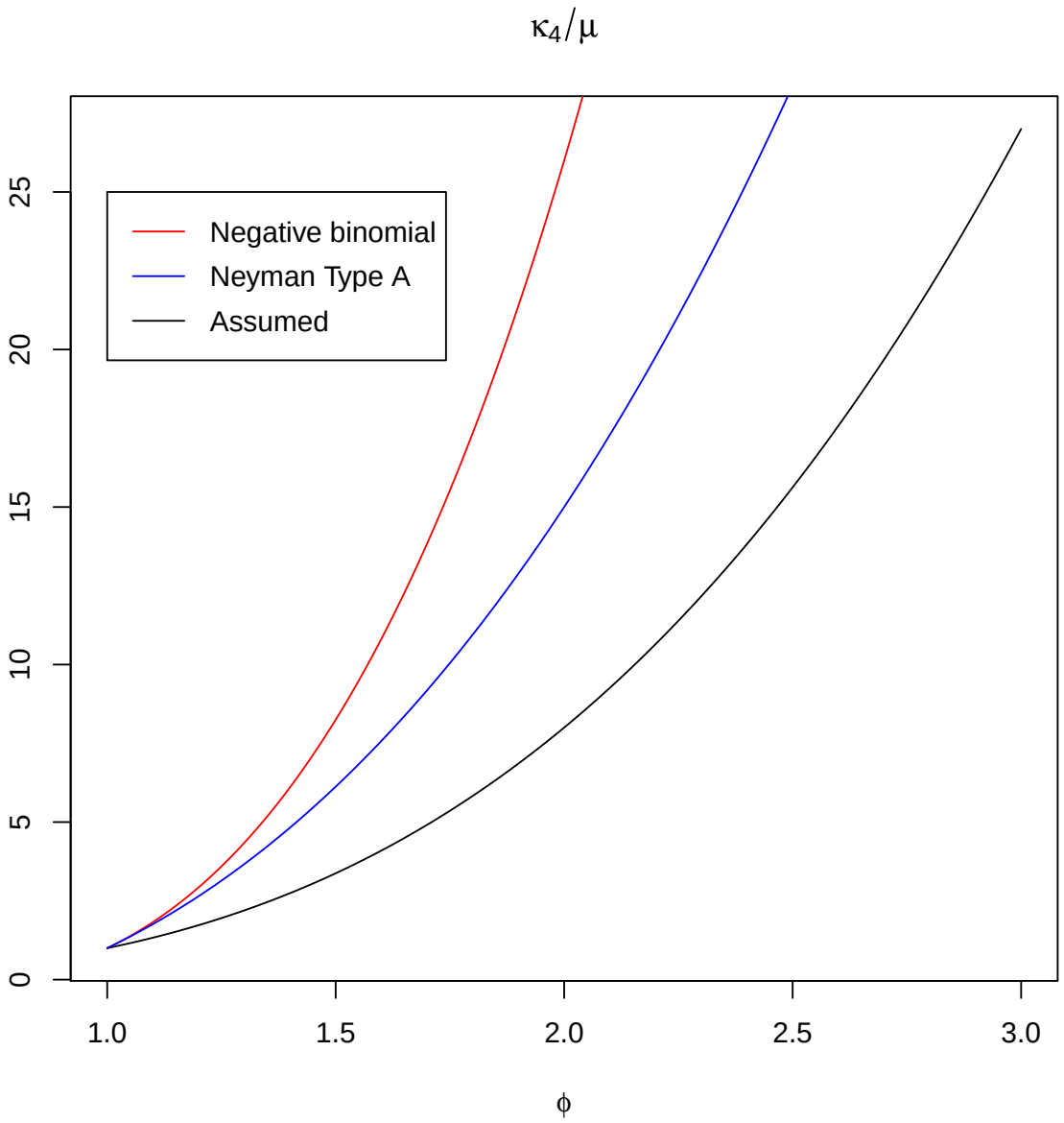
Further Work

- Robustness to assumptions about third and fourth cumluants?
- Asymptotic theory
- Binomial data
- Mark-recapture (multinomial) data
- Quadratic variance-mean relationship?
- Residual diagnostics?
- Bayesian model-checking

Further Work: Robustness to assumptions



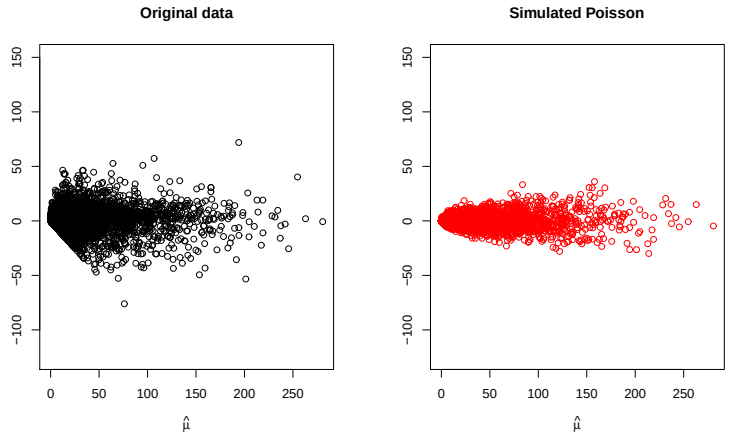
Further Work: Robustness to assumptions

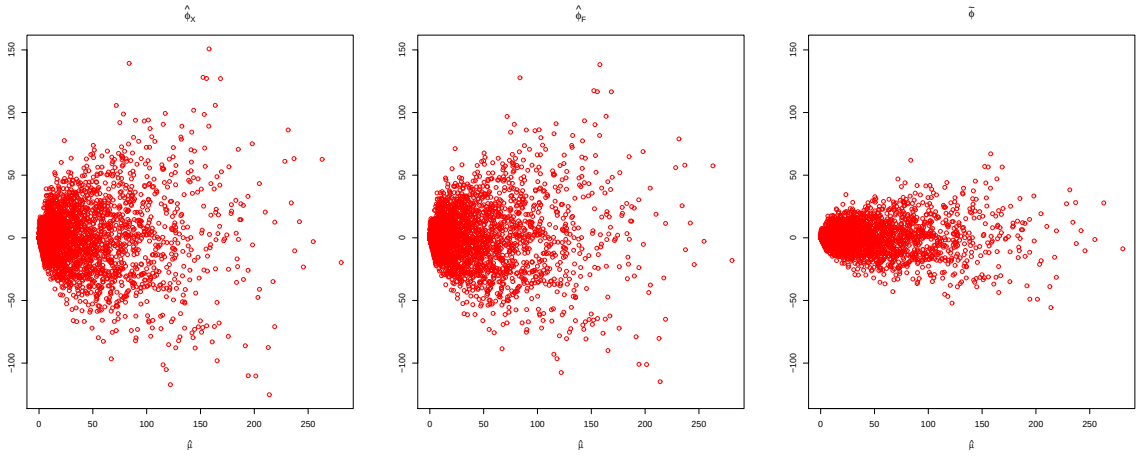


Further Work

- Robustness to assumptions about third and fourth cumluants?
- Asymptotic theory
- Binomial data
- Mark-recapture (multinomial) data
- Quadratic variance-mean relationship?
- Residual diagnostics?
- Bayesian model-checking

Further Work: Residual diagnostics





Further Work

- Robustness to assumptions about third and fourth cumluants?
- Asymptotic theory
- Binomial data
- Mark-recapture (multinomial) data
- Quadratic variance-mean relationship?
- Residual diagnostics?
- Bayesian model-checking