WEIGHTED LIKELIHOOD NEGATIVE BINOMIAL REGRESSION

Michael Amiguet[1], Alfio Marazzi[1], Victor Yohai[2]

1 - University of Lausanne, Institute for Social and Preventive Medicine, Lausanne, Switzerland

2 - University of Buenos Aires, Mathematics Department, Buenos Aires, Argentina

michael.amiguet@chuv.ch

ICORS 2013, Saint Petersburg, 11 July 2013

Outline

1. Approach
2. The model
3. The method
4. Empirical results
5. Example
6. Conclusion and perspectives

1. Approach

The weighted likelihood approach proposed by Agostinelli and Markatou (1998) provides high breakdown point and fully efficient estimators in situations where the errors are i.i.d. variables.
$\rightarrow$ Weights are constructed by comparing the empirical distribution of the residuals to a theoretical distribution

The method we propose allows to apply weighted likelihood estimation in situations where the distribution of the errors is dependent on the covariates, like Poisson regression, or negative binomial regression.

2. The model

We consider the following negative binomial regression framework:
Let $\mathrm{NB}_{\alpha,\mu}$ be the family of negative binomial distributions and $Y_{\alpha,\mu} \sim \mathrm{NB}_{\alpha,\mu}$. Then

- $E(Y_{\alpha,\mu}) = \mu$
- $\mathrm{var}(Y_{\alpha,\mu}) = \mu + \alpha\mu^2$

Regression model:
Response $Y_{\alpha_0,\mu_0(x)} \sim \mathrm{NB}_{\alpha_0,\mu_0(x)}$, where

- x is a covariate vector and $\mu_0(x) = h^{-1}(\beta_0^{\mathsf{T}} x)$
- h is a given link function
- β_0 is a vector of unknown parameters

$\rightarrow$ The errors $Y_{\alpha_0,\mu_0(x)} - \mu_0(x)$ are not i.i.d. and depend on x
$\rightarrow$ They cannot be standardized as in the normal model

We propose a method to estimate α_0 and β_0.

3. The method

The method is a weighted likelihood procedure.

Let $(x_1, y_1), ..., (x_n, y_n)$ be a random sample and use $\theta = (\alpha, \beta)$ and $z_i = (x_i, y_i)$.

We construct weights $w(z_i, \theta)$ and define the estimator of θ as the solution of

$$\sum_{i=1}^{n} w(z_i, \theta) s(\theta, z_i) = 0,$$

where $s(\theta, z)$ is the vector of usual score functions.

Construction of the weights

We define the "tail probabilities" as

$$p_\theta(z_i) = P(Y_{\alpha,\mu(x_i)} \leq y_i) - u_i P(Y_{\alpha,\mu(x_i)} = y_i),$$

where $u_1, \ldots, u_n$ are random numbers generated from the uniform distribution on $[0, 1]$.

Key feature: if $\theta = (\alpha_0, \beta_0)$ then $p_\theta(z_i)$, $i = 1, \ldots, n$ is a sample from a uniform distribution on $[0, 1]$.

Next, we consider the following transformation of the tail probabilities:

Define $q_\theta(z_i)$ as

$$q_\theta(z_i) = \Phi^{-1}(p_\theta(z_i)),$$

where Φ is the standard normal cdf.

If $\theta = (\alpha_0, \beta_0)$, then $q_\theta(z_i)$, $i = 1, \ldots, n$ is a sample from a standard normal distribution.

The weights will be based on a measure of discrepancy between the empirical distribution of the $q_\theta(z_i)$ and the standard normal distribution.

The observations with a large discrepancy will receive small weights.

The weights are defined following a procedure proposed by Agostinelli and Markatou (1998):

Let $\widehat{F}_\theta(.)$ denote the empirical cdf of $q_\theta(z_1), ..., q_\theta(z_n)$; let

$$f^*_\theta(s) = \int k(s,t,h) d\widehat{F}_\theta(t)$$

be a kernel density estimator of the density of $q_\theta(z_i)$, and define $\varphi^*(s)$ as

$$\varphi^*(s) = \int k(s,t,h) d\Phi(t).$$

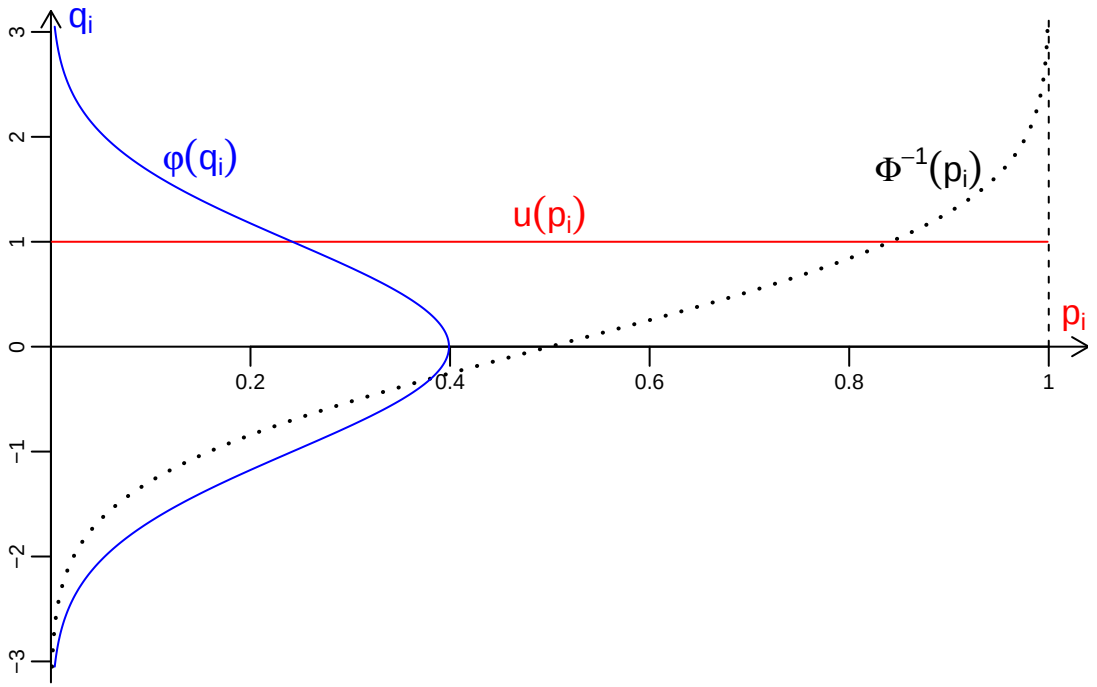
As a local measure of the discrepancy between $f^*_\theta(s)$ and $\varphi^*(s)$, consider, for each observation, its *Pearson residual* $\delta(z_i, \theta)$, defined as

$$\delta(z_i, \theta) = \frac{f^*_\theta(q_\theta(z_i))}{\varphi^*(q_\theta(z_i))} - 1$$

The transformation $q_i = \Phi^{-1}(p_i)$ is important for robustness purposes. To get a high breakdown point and small contamination biases: the less likely an observation under the model, the smaller its weight.

$\rightarrow$ Need an unlikely observation to have large discrepancy i.e. small theoretical density.

Unlikely observations have p_i close to 0 or 1. For such p_i, $u(p_i) = 1$ is not small, however $\varphi(q_i)$ is small.

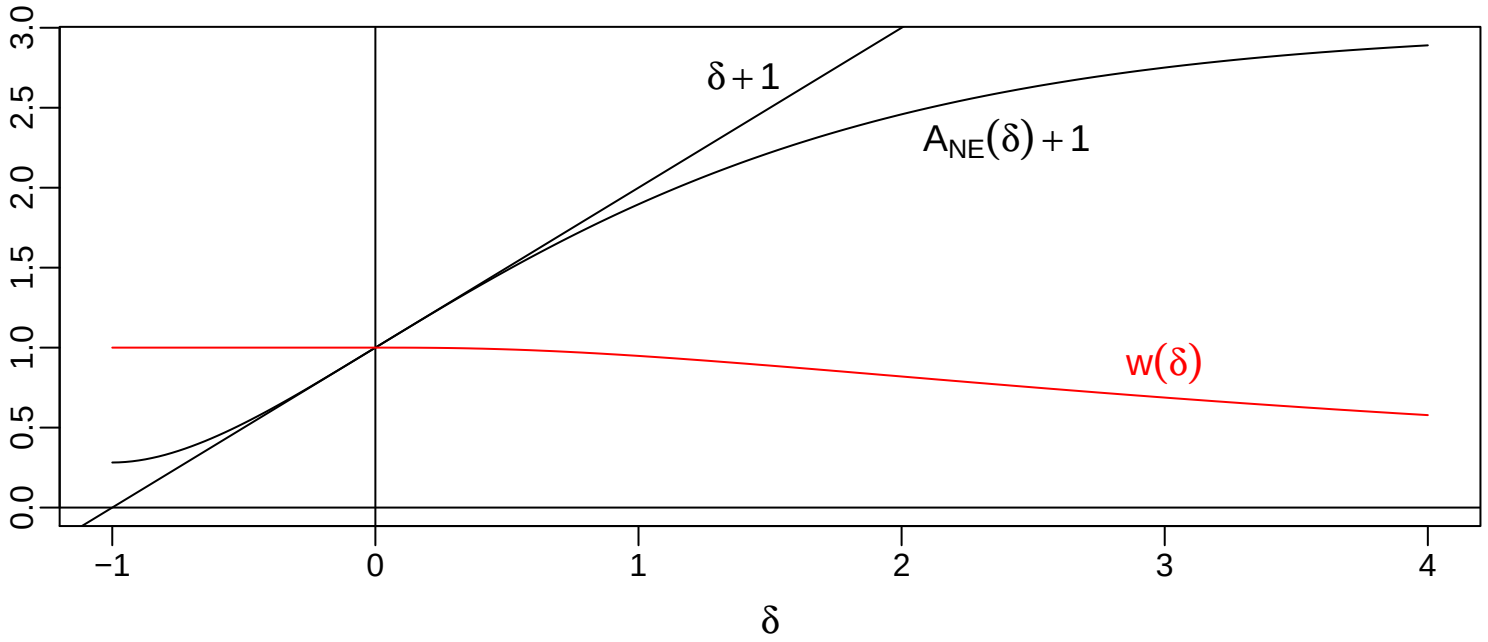


The weights are then defined as

$$w(z_i, \theta) = \min \left\{ 1, \frac{[A(\delta(z_i, \theta)) + 1]^+}{\delta(z_i, \theta) + 1} \right\},$$

where $A(.)$ is a residual adjustment function (Lindsay, 1994), e.g. the negative exponential residual adjustment function

$$A_{\text{NE}}(\delta) = 2 - (2 + \delta) \exp(-\delta).$$



If the model is correct, $\delta(z_i, \theta) = \frac{f^*_\theta(q_\theta(z_i))}{\varphi^*(q_\theta(z_i))} - 1$ converges to 0 and so the weights converge to 1, in which case we recover the maximum likelihood estimator.

This confers high efficiency to the weighted likelihood estimator (WLE), defined through

$$\sum_{i=1}^{n} w(z_i, \theta) s(\theta, z_i) = 0.$$

The initial estimator

The calculation of $\widehat{\theta}$ is done via an iterative algorithm which needs a starting value. In case the estimating equation has multiple roots, we need a robust starting value to avoid convergence to a bad root.

We use a combination of two existing methods:

1. The *maximum rank correlation estimator* (Han, 1987)

   - Maximizes the Kendall correlation $G_n(\beta)$ between the response vector y and the predictor $\beta^{\mathsf{T}}x$:

$$G_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} \{y_i > y_j\}\{\beta^{\mathsf{T}}x_i > \beta^{\mathsf{T}}x_j\}$$

   - Intercept and dispersion parameter α are not identified

   - Slopes are identified up to a scale coefficient

2. An M-Type estimator proposed by V. Yohai

   - Extension to the regression context of the *optimal robust estimate using the Hellinger distance* (Marazzi and Yohai, 2010)

   - Used to estimate the intercept, the dispersion parameter and the scale coefficient on the slopes

Desirable properties of the initial estimator:

- $\sqrt{n}$-consistency facilitates the proof of the asymptotic normality of the WLE.

- High breakdown point: the WLE generally inherits the breakdown point of the initial estimator.

4. Empirical results

Without outliers

We performed simulations with

$$Y_{\alpha_0,\mu_0(x)} \sim \mathsf{NB}_{\alpha_0,\mu_0(x)}; \quad \mu_0(x) = \exp(\beta_0^\mathsf{T} x)$$

in the 2 following models:

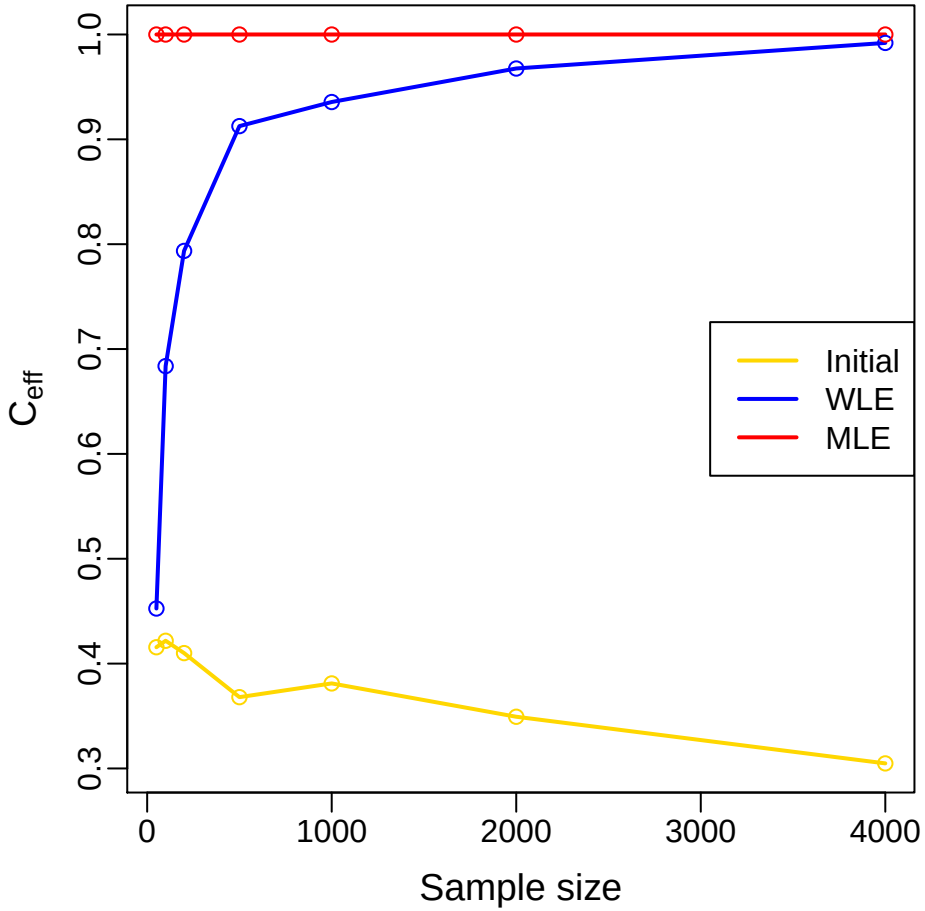
- $\beta_0^\mathsf{T} = (1, 1.5)$; $\alpha_0 = 1.2$
- $\beta_0^\mathsf{T} = (0.5, 0.85, 0.85)$; $\alpha_0 = 0.8$

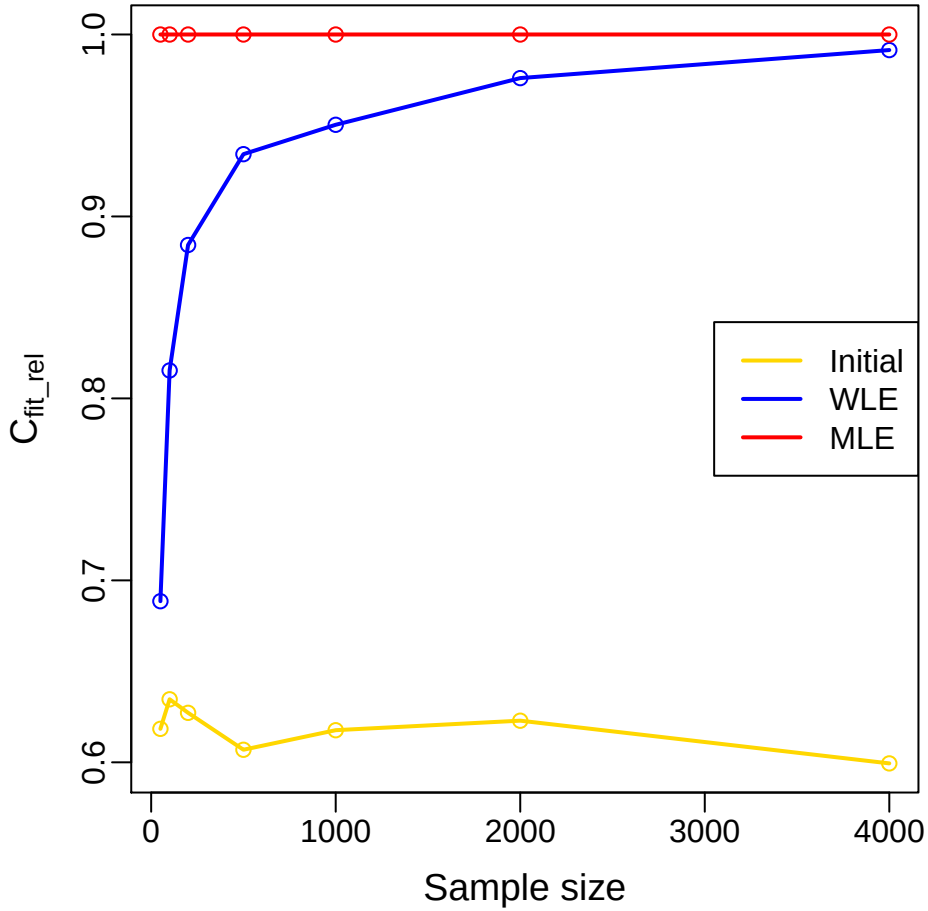
In both models, β_{01} is the intercept.
In each case we performed a simulation with 1000 replications, for sample sizes ranging from 50 to 4000.
We consider two performance criteria:

- Efficiency measure: $C_{\mathsf{eff}} = \dfrac{\sum_j \mathsf{mse}(\beta_j^{MLE}) + \mathsf{mse}(\alpha_j^{MLE})}{\sum_j \mathsf{mse}(\widehat{\beta}_j) + \mathsf{mse}(\widehat{\alpha}_j)}$
- Goodness of fit measure: $C_{\mathsf{fit}} = \mathsf{mean}_{\mathsf{repl}} \left(\mathsf{mean}_i \left(\dfrac{|y_i - \mu_0(x_i)|}{\sqrt{\mu_0(x_i) + \alpha_0 \mu_0(x_i)^2}} \right) \right)$

$$C_{\text{eff}} = \frac{\sum_j \text{mse}(\beta_j^{MLE}) + \text{mse}(\alpha_j^{MLE})}{\sum_j \text{mse}(\hat{\beta}_j) + \text{mse}(\hat{\alpha}_j)}$$

$$C_{\text{fit\_rel}} = \frac{C_{\text{fit}}(MLE)}{C_{\text{fit}}(\text{estimator})}$$

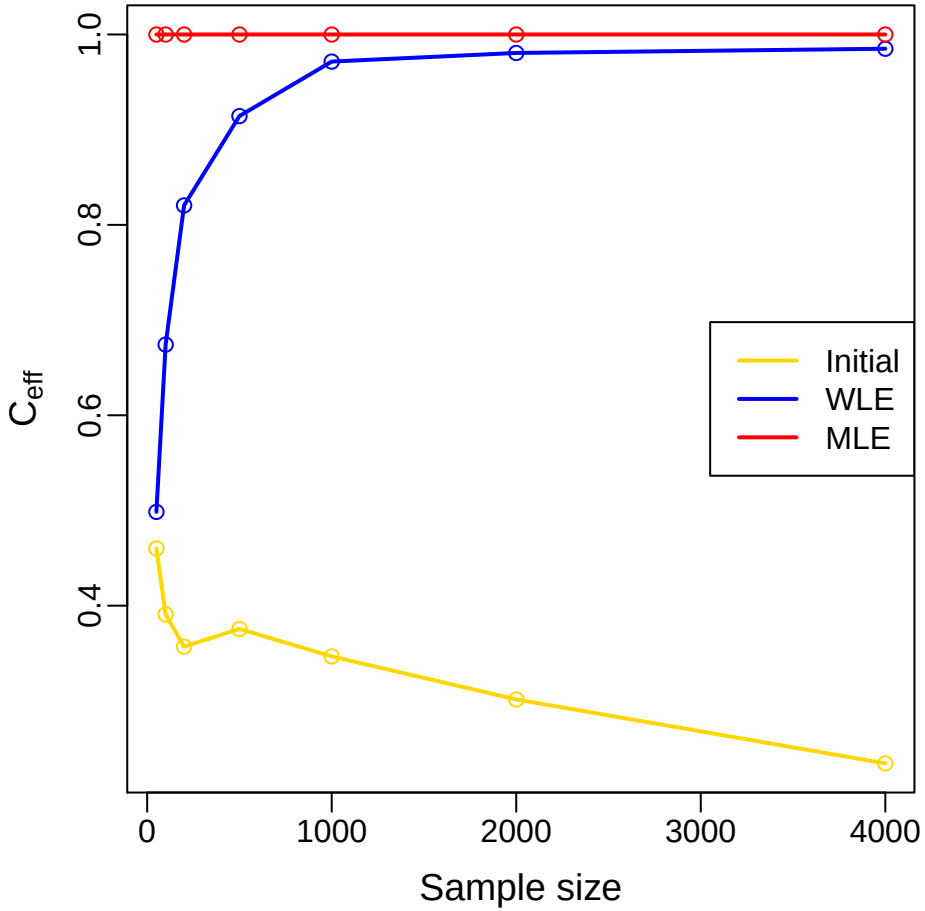


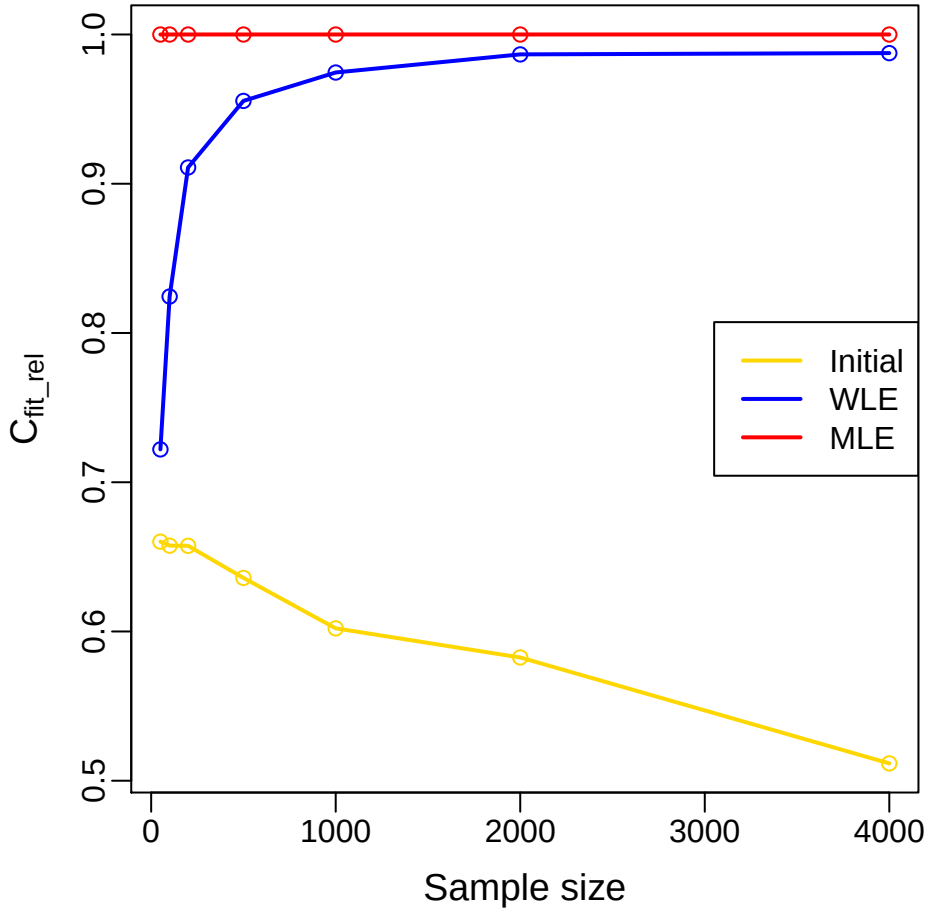


Model 1: $\beta_0^{\mathsf{T}} = (1, 1.5)$; $\alpha_0 = 1.2$

$$C_{\text{eff}} = \frac{\sum_j \text{mse}(\beta_j^{MLE}) + \text{mse}(\alpha_j^{MLE})}{\sum_j \text{mse}(\hat{\beta}_j) + \text{mse}(\hat{\alpha}_j)}$$

$$C_{\text{fit\_rel}} = \frac{C_{\text{fit}}(MLE)}{C_{\text{fit}}(\text{estimator})}$$



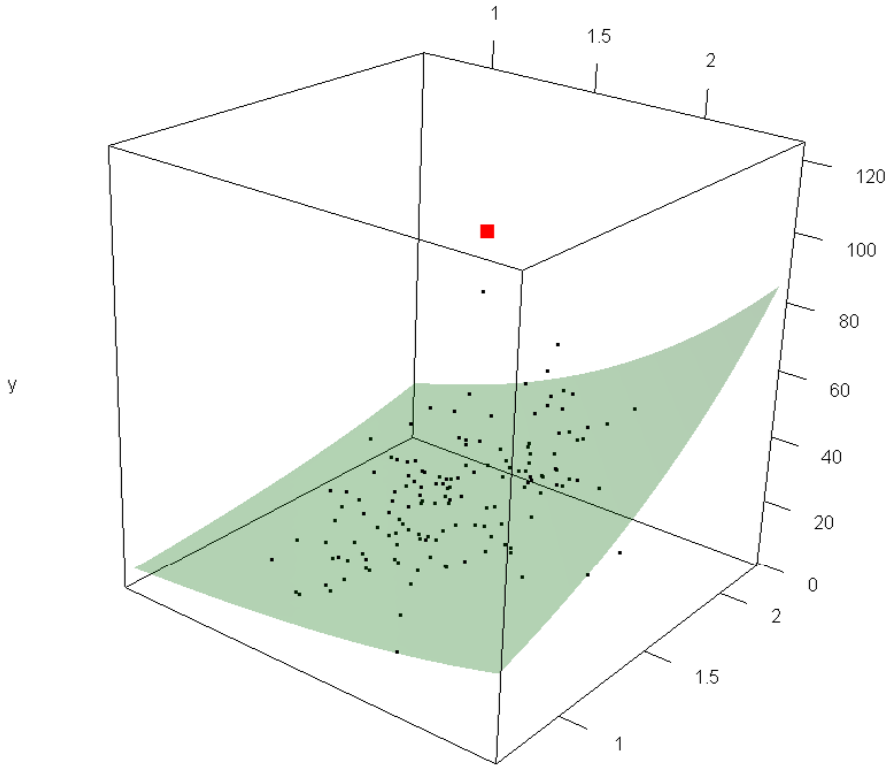


Model 2: $\beta_0^{\mathsf{T}} = (0.5, 0.85, 0.85)$; $\alpha_0 = 0.8$

In the presence of outliers

In order to test the estimators' resistance to outliers, we generated 100 samples of size 150 and replaced an increasing fraction of the observations by outliers. The outliers were placed at the edge of the point cloud with respect to x, and further and further from it in the y direction.

We used the model $\beta_0^T = (0.5, 0.85, 0.85)$; $\alpha_0 = 0.2$.



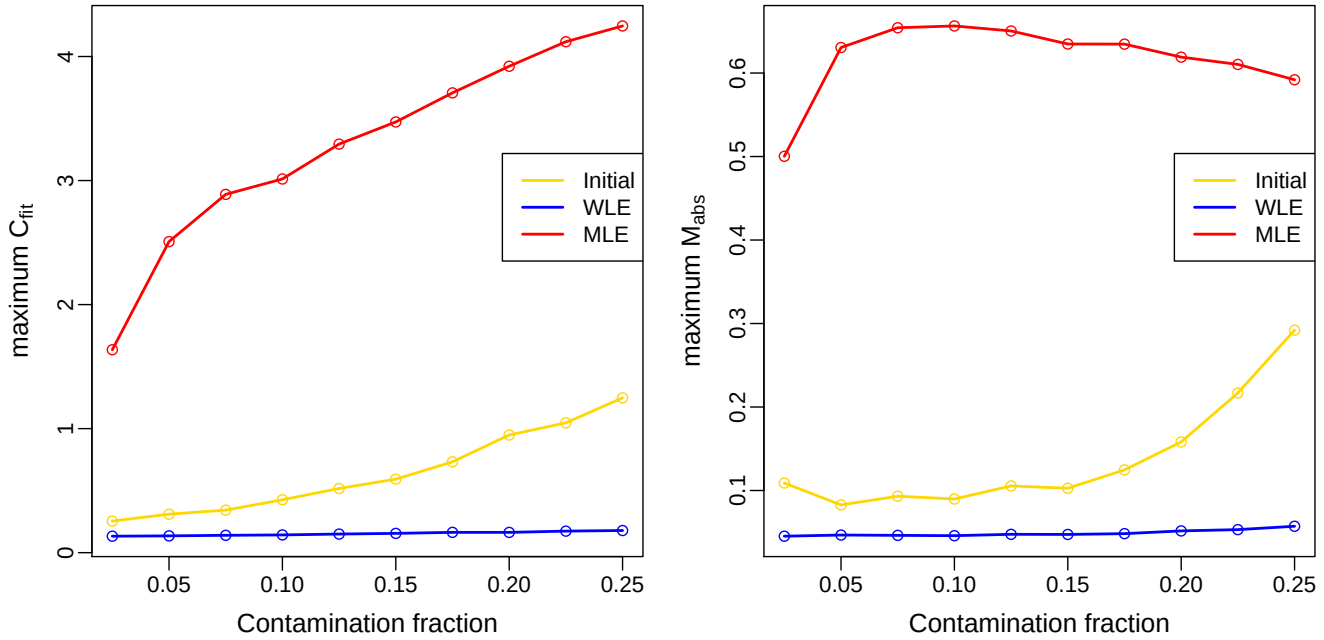
Robustness assessment for each outlier rate and position:

- β's: goodness of fit measure

$$C_{\text{fit}} = \text{mean}_{\text{repl}} \left(\text{mean}_i \left(\frac{|\widehat{y}_i - \mu_0(x_i)|}{\sqrt{\mu_0(x_i) + \alpha_0 \mu_0(x_i)^2}} \right) \right)$$

- α: mean absolute error $M_{\text{abs}} = \text{mean}_{\text{repl}}(|\widehat{\alpha} - \alpha_0|)$

The graphs show the largest C_{fit} and M_{abs} obtained for each contamination fraction (over outlier position).



5. Example

We consider hospital length of stay (LOS) data in the state of Lausanne, Switzerland.

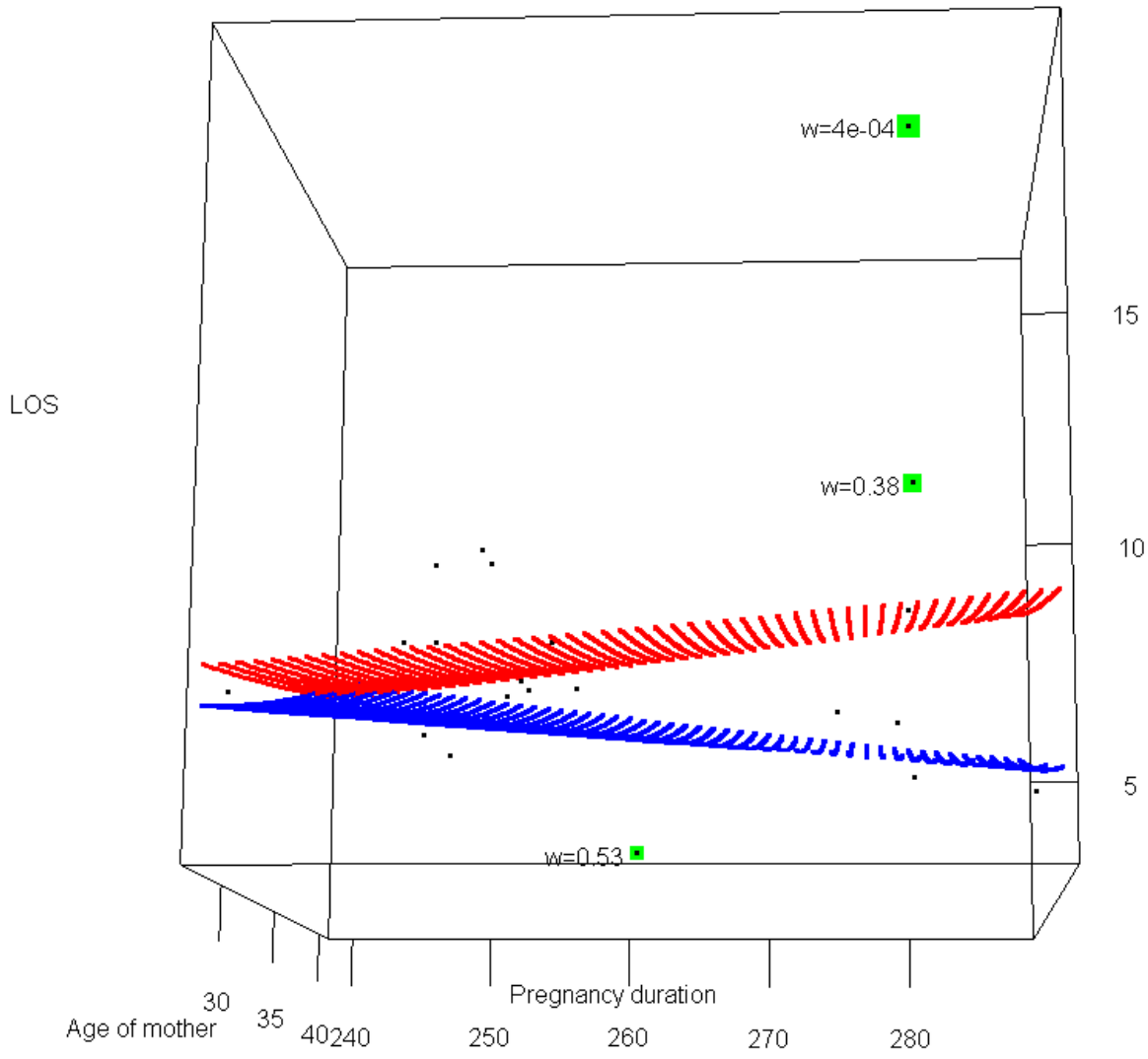
23 stays in 2010 for neonates classified into Diagnosis Related Group entitled "Neonate, birth weight >2499g, without significant operating room procedures, with other problems".

We model the LOS with two independent variables:

- Age of mother
- Pregnancy duration

Younger mothers and longer pregnancy durations are known to imply shorter LOS.

Result:



Red surface: MLE Blue surface: WLE

6. Conclusion and perspectives

- We propose a new robust and efficient estimation method for negative binomial regression
- The simulation results are promising, showing high robustness and efficiency performances
- Consistency, efficiency and robustness theory are being developed
- The central idea is to use "tail probabilities" in order to get i.i.d. residuals to which the weighted likelihood method can be applied
- This idea could be applied to a large variety of regression frameworks where the errors are not necessarily i.i.d. and can involve shape parameters

References

Agostinelli, C., Markatou, M., 1998, A one-step robust estimator for regression based on the weighted likelihood reweighting scheme. Statistics & Probability Letters 37, 342-350.

Lindsay, B.G., 1994, Efficiency versus robustness: The case for minimum Hellinger distance and related methods. Ann. Statist. 22, 1018-1114.

Han, A.K., 1987, Non-parametric analysis of a generalized regression model. Journal of Econometrics 35, 303-316.

Marazzi, A., Yohai, V.J., 2010, Optimal robust estimates using the Hellinger distance. Advances in Data Analysis and Classification, 1-11.