Robust estimators for generalized linear models with a dispersion parameter

Michael Amiguet\textsuperscript{(1)}, Alfio Marazzi\textsuperscript{(1)}, Marina Valdora\textsuperscript{(2)}
and Victor Yohai\textsuperscript{(2)},

\textsuperscript{(1)} Université de Lausanne, \textsuperscript{(2)} Universidad de Buenos Aires

March 29, 2017

Abstract

Highly robust and efficient estimators for the generalized linear model with a dispersion parameter are proposed. The estimators are based on three steps. In the first step the maximum rank correlation estimator is used to consistently estimate the slopes up to a scale factor. In the second step, the scale factor, the intercept, and the dispersion parameter are consistently estimated using a MT-estimator of a simple regression model. The combined estimator is highly robust but inefficient. Then, randomized quantile residuals based on the initial estimators are used to detect outliers to be rejected and to define a set \( S \) of observations to be retained. Finally, a conditional maximum likelihood (CML) estimator given the observations in \( S \) is computed. We show that, under the model, \( S \) tends to the complete sample for increasing sample size. Therefore, the CML tends to the unconditional maximum likelihood estimator. It is therefore highly efficient, while maintaining the high degree of robustness of the initial estimator. The case of the negative binomial regression model is studied in detail.
1 Introduction

In recent years, several extensions of the generalized linear models (GLM; Nelder and Wedderburn, 1972) have been proposed to increase flexibility in modelling complex data structures. We consider the case where the response distribution does not necessarily belong to the exponential family and where a dispersion parameter is present. For this case, we will propose highly efficient and highly robust estimators. We focus on the Negative Binomial (NB) regression model, but we also consider the Beta regression model as an example with continuous response. NB regression (see Hilbe, 2008) extends Poisson regression for modeling count data in presence of overdispersion. Beta regression (Ferrari and Cribari-Neto, 2004) is a tool for modelling continuous responses which are restricted to the interval $[0, 1]$, such as rates and proportions. Both these models have important biometrical applications. NB regression is the most popular model for the analysis of hospital length of stay (e.g., Austin et al., 2002; Hilbe, 2008; Carter and Potts, 2014). Among other applications, we also mention its use to model falls data (Aeberhard et al., 2014). Applications of Beta regression in medicine can be found in Hunger et al. (2011), Swearingen et al. (2011), and Seow et al. (2012) among others.

Usually, the parameters are estimated by means of the maximum likelihood (ML) principle, which provides fully efficient estimators when the observations follow the model. ML procedures to fit the NB regression have been implemented in popular statistical software such as STATA, SAS, SPSS, and in the R package MASS (Venables and Ripley, 1999). An implementation of the Beta regressions can be found in R (Cribari–Neto and Zeiles, 2010).

Unfortunately, the ML estimator is extremely sensitive to the presence of outliers in the sample, i.e., observations with unexpectedly extreme values in the response variable. This sensitivity increases when these extreme responses come together with large values in the covariate space. In certain applications, such as the analysis of
hospital length of stay, the proportion of outliers - often called the contamination level - may be as high as 10%. Such a level of contamination can not only strongly bias the coefficient estimates but also lead to overestimating the dispersion parameter. As a consequence, inferences based on the ML fit may be badly misguided.

There are two basic approaches to detect outliers and assess their influence. The first one makes use of diagnostic tools based on ML residuals. Specific proposals for GLM are described by Davison and Snell (1991) and proposals for Beta regression by Espinheira et al. (2008) and Rocha and Simas (2010). However, this strategy may fail because the ML estimators may be distorted and residuals corresponding to outliers are not necessarily large and visible; a well known “masking effect” is described in Maronna et al. (2006, p. 179). A better strategy, is the use of a robust estimator, that is an estimator which is not very sensitive to the presence of outliers. There are many proposals of robust estimators for GLM models (e.g., Künsch et al. 1989; Cantoni and Ronchetti, 2001). However, most of them do not admit an extra parameter besides the coefficient vector. A few robust estimators of the parameters of the NB distribution in the absence of covariates have been considered in Cadigan and Chen (2001) and Amiguet (2011). Marazzi and Yohai (2010) implemented M estimators satisfying Hampel’s optimality principle (Hampel et al., 1986) for multiparameter families of distributions including NB and Beta. Yet, it is cumbersome to extend these estimators to the regression case. Aeberhard et al. (2014) proposed a generalized M (GM) estimator for NB regression. Unfortunately, GM estimators have several drawbacks. In particular, their degree of robustness - as measured by the breakdown point - decreases when the number of covariates increases (Maronna et al., 2006, p.149). Moreover, GM estimators depend on “tuning constants” that are chosen to attain a given level of efficiency at a specified model with known parameter values; but the parameters are unknown before estimation. In order to ensure consistency at the (unknown) model, several corrections have to be implemented adding complexity to
the computation and increasing the computing time. At present, no robust procedure for Beta regression has been published.

In this paper, we introduce a novel class of estimators for GLMs with a dispersion parameter. Following an approach that we have developed in previous papers for different models (Marazzi and Yohai, 2004; Locatelli, Marazzi, Yohai, 2010), we consider a three phase procedure. In the first phase, a highly robust but possibly inefficient estimator is computed. This initial estimator allows outlier identification. Finally, a conditional ML procedure is used, where observations are constrained to belong to a subsample free of large outliers. However, in the absence of outliers, this subsample tends to the original sample if its size increases and, therefore, the final estimator is asymptotically fully efficient. Nevertheless, it maintains a similar degree of robustness as the initial estimator. Conditional ML estimators have also been used by Cuesta-Albertos, Matrán, and Mayo-Iscar (2008) to define multivariate robust location and dispersion estimators.

In Section 2 we introduce the general model. Section 3 defines the estimators. The efficiency and the robustness of the new procedures are demonstrated in Section 4 by means of Monte Carlo experiments. Two examples, where the procedures are applied to hospital length of stay data are described in Section 5. The discussion in Section 6 ends the paper. Three appendices provide proofs and some supplementary material. The methods we are proposing in this paper have been implemented in the R package “robustGLM” available on the Comprehensive R Archive Network.

2 The model

Let $F_{\mu,\alpha}(y)$ denote a general family of discrete or continuous distribution functions, where $\mu$ is the mean and $\alpha$ is a dispersion parameter, and let $f_{\mu,\alpha}(y)$ denote the corresponding probability (density) function. We will focus on two specific examples
of families, one discrete and one continuous:

- the NB family:
  \[
  f_{\mu,\alpha}(y) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(1/\alpha)\Gamma(\mu + y + 1)} (\frac{\alpha\mu}{\alpha\mu + 1})^y, \quad y = 0, 1, 2, ..., \alpha \geq 0, \mu \geq 0; \quad (1)
  \]

- the Beta family:
  \[
  f_{\mu,\alpha}(y) = \frac{\Gamma(1/\alpha)}{\Gamma(\mu/\alpha)\Gamma((1-\mu)/\alpha)} y^{\mu/\alpha - 1}(1-y)^{(1-\mu)/\alpha - 1}, \quad 0 < y < 1, \alpha \geq 0, \mu \geq 0. \quad (2)
  \]

In both cases, the parametrization has been chosen so that the expected value is \(\mu\).

In the NB case, the variance is \(\mu + \alpha\mu^2\); in the Beta case, the variance is \(\mu(1-\mu)/(1+1/\alpha)\). In both cases, fixing \(\mu\), the variance increases with \(\alpha\).

We will need the following assumption on \(F_{\mu,\alpha}(y)\), which is satisfied in our examples:

**Assumption A:** For any \(\alpha\), \(Y_1 \sim F_{\mu_1,\alpha}(y)\), \(Y_2 \sim F_{\mu_2,\alpha}(y)\), if \(\mu_2 > \mu_1\) then \(Y_2 \succ Y_1\), where “\(\succ\)” means “stochastically larger”.

Suppose that a response \(Y\) and a vector \(X = (X_1, ...X_p)^T\) of covariates are observed. We consider the following class of regression models

\[
Y \mid X = x \sim F_{h(x^T\beta_0),\alpha_0},
\]

where \(h\) is a strictly increasing known link function, and \(\beta_0 = (\beta_{01}, ...\beta_{0p})^T\) is a vector of coefficients. We assume that \(X_1\) is constantly equal to one, that is, \(\beta_{01}\) is an intercept. We will use the notations \(x^T = (1, x^*T)\), \(\beta_0^* = (\beta_{01}, \beta_0^*T)\), \(\gamma_0 = \beta_0^*/||\beta_0^*||\), and \(\theta = (\beta, \alpha)\).

We assume that a random sample \((x_1, y_1), ..., (x_n, y_n)\) is available. The ML estimator of \(\theta_0 = (\beta_0, \alpha_0)\) maximizes the log-likelihood of the sample given by

\[
\mathcal{L}(\theta) = \sum_{i=1}^{n} \ln \left(f_{h(x_1^T\beta),\alpha}(y_i)\right).
\]

The ML estimator is very efficient but not robust. We want to obtain highly robust and efficient estimators of \(\beta_0\) and \(\alpha_0\).
3 Estimation procedure

The proposed procedure starts with the computation of a very robust but not necessarily efficient initial estimator which provides the tool for outlier identification. Then, a conditional ML approach is used - where the outliers are removed - which provides a fully efficient estimator.

Most familiar highly robust estimators of regression, such as LMS, LTS, and S estimators (see, e.g., Maronna et al., 2006), are based on the minimization of a robust measure of the residual scale, such as an M scale (Huber, 1980). These estimators have been used as initial estimators of well known highly robust and efficient procedures, such as MM (Yohai, 1987), and TML (Marazzi and Yohai, 2004) estimators. However, for the regression models we are considering here, a different approach has to be used because the residual distribution may depend on the covariates and residual measures of scale are not available in this case. We therefore propose an approach based on the maximum rank correlation (MRC) estimator introduced by Han (1987a) and Han (1987b). However, the MRC estimator identifies the scaled slopes $\gamma_0 = \beta_0^*/||\beta_0^*||$, but it does not identify the intercept $\beta_{01}$, the dispersion parameter $\alpha_0$, and the scale factor $\eta_0 = ||\beta_0^*||$. So, we need to estimate these three parameters separately. The complete proposal can then be summarized as follows:

**Step 1** Compute the MRC estimator $\tilde{\gamma}$ of $\gamma_0$. In addition, compute robust and consistent estimators $\tilde{\beta}_1$, $\tilde{\alpha}$, and $\tilde{\eta}$, of $\beta_{01}$, $\alpha_0$, and $\eta_0$. Then, initial estimators of $\beta_0$ and $\alpha_0$ are given by $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\eta}_0)^T$ and $\tilde{\alpha}$ respectively.

**Step 2** Compute randomized quantile residuals $z_i$ (Dunn and Smyth, 1996) based on the initial model and use them to define cutoff values $\tilde{a}$ and $\tilde{b}$, so that influential outliers are defined as observations such that $z_i \notin [\tilde{a}, \tilde{b}]$.

**Step 3** Compute a conditional ML estimator of $\theta_0$ given $z_i \in [\tilde{a}, \tilde{b}]$. 
In the following subsections, we provide a detailed description of each single step.

3.1 The initial estimator

For a given coefficient vector \( \gamma = (\gamma_2, \ldots, \gamma_p)^T \), the Kendall’s \( \tau \) correlation coefficient between the responses \( y_i \)'s and the linear combinations \( \gamma^T x_i^* \)'s is given by

\[
\tau(\gamma) = \frac{1}{n(n-1)} \sum_{i \neq j} I[(\gamma^T x_j^* - \gamma^T x_i^*)(y_j - y_i) \geq 0]
\]

and the maximum rank correlation (MRC) estimator of \( \gamma_0 \) is defined by

\[
\tilde{\gamma} = \arg \min_{\|\gamma\| = 1} \tau(\gamma).
\]

The robustness of Kendall’s \( \tau \) correlation coefficient has been studied by Alfons et al. (2016). Under the assumption A, the MRC estimator strongly converges to \( \gamma_0 \) for any strictly increasing \( h \) (Han, 1987a); it is also root \( n \) consistent and asymptotically normal (Sherman, 1993).

To compute the MRC estimator one can utilize a subsampling procedure. Note that the simple evaluation of the objective function requires \( O(n^2) \) calculations, but an algorithm using \( O(n \log n) \) calculations has been proposed by Abrevaya (1999). However, in the Monte Carlo experiments described in Section 4, we used the very fast function maxCorGrid of the R package ccaPP (Alfons, 2015) based on an alternate grid algorithm described in Alfons et al. (2016).

We now turn to the estimation of \( \beta_{01}, \alpha_0, \) and \( \eta_0 \), necessary to complete the initial estimator. We observe that \( h(x^T \beta_0) = h(\beta_{01} + \eta_0^T x^*) \). Since \( \tilde{\gamma} \) is close to \( \gamma_0 \), we approximate \( \gamma_0^T x_i^* \) by \( v_i = \tilde{\gamma}^T x_i^* \) and consider the simple regression model with just one covariate:

\[
Y \mid v \sim F_{h(\beta_{01} + \eta_0 v), \alpha_0}.
\]

For this model and a given value \( \alpha \) of the unknown \( \alpha_0 \), we have many highly robust estimators \( \tilde{\beta}_{01}(\alpha), \tilde{\eta}(\alpha) \), of \( \beta_{01} \) and \( \eta_0 \). Examples are: the conditionally unbiased
bounded influence estimator of Künsch et al. (1989), the RQL estimator of Cantoni and Ronchetti (2001), and the weighted MT estimators of Valdora and Yohai (2014). Finally, to estimate \( \alpha_0 \), we consider a bounded function \( \psi(y, \mu, \alpha) \) such that, for all \( \mu \), we have

\[
E_{\mu, \alpha}[\psi(y, \mu, \alpha)] = 0. \tag{6}
\]

Then, for any fixed \( \mu \), if \( y_1, y_2, \ldots y_n \) is a random sample of \( NB(\mu, \alpha_0) \), the M estimator of \( \alpha \) satisfying the equation

\[
\sum_{i=1}^n \psi(y_i, \mu, \alpha) = 0
\]

is Fisher consistent for \( \alpha_0 \). Then, an initial consistent estimator \( \tilde{\alpha} \) of \( \alpha_0 \) is obtained by solving

\[
\sum_i \psi(y_i, h(\tilde{\beta}_1^*(\alpha) + \tilde{\eta}^*(\alpha)v_i), \alpha) = 0. \tag{7}
\]

The Fisher consistency of this estimator is immediate. In fact, asymptotically, \( h(\tilde{\beta}_1^*(\alpha) + \tilde{\eta}^*(\alpha)v_i) = E(y_i) \), and then by (6)

\[
E(\psi(y, \mu_i, \alpha)) = 0.
\]

Once \( \tilde{\alpha} \) is computed, we define the initial estimators of \( \beta_{01} \) and \( \eta_0 \) by \( \bar{\beta}_1 = \tilde{\beta}_1^*(\tilde{\alpha}) \), \( \bar{\eta} = \tilde{\eta}^*(\tilde{\alpha}) \). In this way we obtain the initial estimators \( \tilde{\beta} = (\bar{\beta}_1, \bar{\eta}) \) of \( \beta_0 \) and \( \tilde{\alpha} \) of \( \alpha_0 \).

We will assume that:

**Assumption B**: \( n^{1/2}(\tilde{\beta} - \beta_0) = O_p(1) \) and \( n^{1/2}(\tilde{\alpha} - \alpha_0) = O_p(1) \).

In the simulations of Section 4 and the examples in Section 5, we use a weighted MT estimator for \( \tilde{\beta}_1^*(\alpha) \), \( \tilde{\eta}^*(\alpha) \) (see appendix 8) and the score function \( \psi \) of the optimal bounded influence estimator according to Hampel (1972) described in Marazzi and Yohai (2010). It can be proved that, under general conditions, the resulting initial estimators \( \tilde{\beta} \) and \( \tilde{\alpha} \) satisfy the assumption B.
3.2 Adaptive cutoff values and outlier detection

We now assume that some preliminary estimator $\tilde{\theta} = (\tilde{\beta}, \tilde{\alpha})$ of $\theta_0$ is available, for example the estimators defined in the previous section. Since the residual distribution depends on the covariates, residuals cannot be used in the usual way for the purpose of highlighting outliers. Instead, we use the randomized quantile residuals (RQR) that were proposed in Dunn and Smyth (1996) for exploratory purposes. Let $\tilde{\mu}_x = h(x^T \tilde{\beta})$. Then, the RQRs are defined by

$$z_i = F_{\tilde{\mu}_x, \tilde{\alpha}}(y_i)$$

in the continuous case and by

$$z_i = F_{\tilde{\mu}_x, \tilde{\alpha}}(y_i) - u_i f_{\tilde{\mu}_x, \tilde{\alpha}}(y_i)$$

in the discrete case, where $\{u_1, ..., u_n\}$ is a sample from a uniform distribution $U[0, 1]$ independent of the original sample $(x_1, y_1), ..., (x_n, y_n)$.

If $\tilde{\theta} = \theta_0$, $\{z_1, ..., z_n\}$ is a sample from $U[0, 1]$. Then, a fixed lower cutoff value $a$ and a fixed upper cutoff value $b$ for the RQRs are simply given by a low, respectively a large quantile of $U[0, 1]$ – e.g., $a = 0.05$ and $b = 0.95$ – and observations such that $z_i \notin [a, b]$ may be identified as outliers. However, we propose the use of “adaptive” cutoff values $\tilde{a}$ and $\tilde{b}$ that, under the assumed model, tend to 0 and 1 respectively, when $\tilde{\beta}$ and $\tilde{\alpha}$ are consistent estimators. Therefore, under the model, i.e., in the absence of outliers, the fraction of observations that are erroneously identified as outliers tends to 0 when the sample size $n \to \infty$.

To define the adaptive cutoff values, we follow a procedure similar to the ones described in Marazzi and Yohai (2004, Section 3.2) and in Gervini and Yohai (2002). Let $F_n$ denote the empirical cdf of $z_1, ..., z_n$ and $F_{n,t}^R$ and $F_{n,t}^L$ be the right and the left truncated versions of $F_n$ for a given $t$ respectively, i.e.,

$$F_{n,t}^R(z) = \begin{cases} 
F_n(z)/F_n(t) & \text{if } z \leq t, \\
1 & \text{otherwise,}
\end{cases}$$
\[ F_{n,t}^L(z) = \begin{cases} \frac{(F_n(z) - F_n(t))/(1 - F_n(t))}{1 - F_n(t)} & \text{if } z \geq t, \\ 0 & \text{otherwise.} \end{cases} \]

We then compare the rights tails of \( F_n \) and the \( U[0, 1] \), looking for the largest \( t \) such that \( F_{n,t}^R(z) \geq z \) for all \( z \geq \zeta_2 \) where \( \zeta_2 \) is a value close to one. More precisely, we define an upper cutoff value as

\[ \tilde{b} = \sup \left\{ t : \inf_{z \geq \zeta_2}(F_{n,t}^R(z) - z) \geq 0 \right\}. \]

In a similar way, we define a lower cutoff value as

\[ \tilde{a} = \inf \left\{ t : \sup_{z \leq \zeta_1}(F_{n,t}^L(z) - z) \leq 0 \right\}, \]

where \( \zeta_1 \) is close to zero.

We assume that:

**Assumption C:** The density \( f(y, \mu, \alpha) \) has a bounded derivative with respect to \( \mu \) and \( \alpha \).

Then, we have the following Theorem, proved in Appendix 9.

**Theorem 1** Assume \( B \) and \( C \). Then

\[ n^{1/2}\tilde{a} = O_p(1), \quad n^{1/2}(\tilde{b} - 1) = O_p(1). \]

Usually, a quite high value of \( \zeta_2 \) is chosen. Our usual choice is \( \zeta_2 = 0.95 \); however, in the presence of a large proportion of high outliers, it may be convenient to use a lower value, e.g., \( \zeta_2 = 0.90 \). Similar considerations apply to the choice of the lower cutoff \( \tilde{a} \) and we usually set \( \zeta_1 = 0.05 \), but \( \zeta_1 = 0.10 \) would allow removing a larger fraction of small observations, such as “excess zeros” in the NB case. (In fact, a very small \( \zeta_1 \) could fail to identify many “excess zeros”, because each one of them corresponds to several distinct \( z_i \)’s and may not emerge as an extremely small value.)
3.3 Final estimator

In the final step, we improve the efficiency of the initial estimator using a conditional ML approach. Suppose first that fixed cutoff values $a$ and $b$ are given and the RQRs are computed. Let $p_{\beta,\alpha}(y \mid x, Z \in [a, b])$ denote the conditional density of $Y$ given $X = x$ and $Z \in [a, b]$, where $Z \sim U[0, 1]$ represents the RQR. Then, the conditional density is of the form

$$p_{\beta,\alpha}(y \mid x, Z \in [a, b]) = f_{h(x^T\beta),\alpha}(y)W_{a,b}(x, \beta, \alpha).$$

(8)

In the continuous case we have

$$W_{a,b}(x, \beta, \alpha) = I(F_{\mu_x,\alpha}^{-1}(a) \leq y \leq F_{\mu_x,\alpha}^{-1}(b)) \frac{b - a}{b - a}.$$

In the discrete case, the following expression (9) for $W_{a,b}(x, \beta, \alpha)$ is derived in the appendix 7. Let, for any $c$,

$$y_x^*(c) = \max\{y : F_{\mu_x,\alpha}(y) \leq c\},$$

and

$$t_{c,x} = \frac{F_{\mu_x,\alpha}(y_x^*(c) + 1) - c}{f_{\mu_x,\alpha}(y_x^*(c) + 1)},$$

Put

$$T_{a,x} = y_x^*(a) + 2, \quad T_{b,x} = y_x^*(b),$$

$$A_x = \{y : T_{a,x} \leq y \leq T_{b,x}\},$$

and

$$Q(x, \beta, \alpha) = F_{\mu_x,\alpha}(T_{b,x}) - F_{\mu_x,\alpha}(T_{a,x} - 1) + f_{\mu_x,\alpha}(T_{a,x} - 1)t_{a,x} + f_{\mu_x,\alpha}(T_{b,x} + 1)(1 - t_{b,x}).$$

Then

$$W_{a,b}(x, \beta, \alpha) = \begin{cases} \frac{1}{Q(x, \beta, \alpha)} & \text{if } y \in A_x, \\ \frac{t_{a,x}}{Q(x, \beta, \alpha)} & \text{if } y = T_{a,x} - 1, \\ \frac{1 - t_{b,x}}{Q(x, \beta, \alpha)} & \text{if } y = T_{b,x} + 1, \\ 0 & \text{if elsewhere.} \end{cases}$$

(9)
We now suppose that $\tilde{a}$ and $\tilde{b}$ are the adaptive cutoff values defined above, and consider the adaptive conditional likelihood function

$$L_{CML}(\theta) = \sum_{i=1}^{n} I(\tilde{a} \leq z_i \leq \tilde{b}) \ln \left( p_{\beta,\alpha}(y_i \mid x_i, z_i \in [\tilde{a}, \tilde{b}]) \right).$$

The conditional maximum likelihood (CML) estimator $\hat{\theta}_{CML} = (\hat{\beta}_{CML}, \hat{\alpha}_{CML})$ is defined by

$$\hat{\theta}_{CML} = \arg \max_{\theta} L_{CML}(\theta).$$

In the discrete case, a slight modification of this definition is convenient. We note that (see appendix 7):

$$\{a \leq z_i \leq b\} = \{T_{a,x_i} \leq y_i \leq T_{b,x_i}\} \cup \{y_i = T^*_{i,x}, u_i \leq t_{a,x_i}\} \cup \{y_i = T^*_{u,x}, u_i \geq t_{b,x_i}\},$$

where $T^*_{i,x} = T_{a,x} - 1$, and $T^*_{u,x} = T_{b,x} + 1$. Then,

$$L_{CML}(\theta) = \sum_{T_{a,x_i} \leq y_i \leq T_{b,x_i}} \ln \left( p_{\beta,\alpha}(y_i \mid x_i, z_i \in [\tilde{a}, \tilde{b}]) \right)$$

$$+ \sum_{y_i = T^*_{i,x}} I(u_i \leq t_{a,x_i}) \ln \left( p_{\beta,\alpha}(y_i \mid x_i, z_i \in [\tilde{a}, \tilde{b}]) \right)$$

$$+ \sum_{y_i = T^*_{u,x}} I(u_i \geq t_{b,x_i}) \ln \left( p_{\beta,\alpha}(y_i \mid x_i, z_i \in [\tilde{a}, \tilde{b}]) \right).$$

Since the $u_i$-s are non-informative, we replace $I(u_i \leq t_{a,x_i})$ and $I(u_i \geq t_{b,x_i})$ by their expected values, and define

$$L_{MCML}(\theta) = \sum_{T_{a,x_i} \leq y_i \leq T_{b,x_i}} \ln \left( p_{\beta,\alpha}(y_i \mid x_i, z_i \in [\tilde{a}, \tilde{b}]) \right)$$

$$+ \sum_{y_i = T^*_{i,x}} t_{a,x_i} \ln \left( p_{\beta,\alpha}(y_i \mid x_i, z_i \in [\tilde{a}, \tilde{b}]) \right)$$

$$+ \sum_{y_i = T^*_{u,x}} (1 - t_{b,x_i}) \ln \left( p_{\beta,\alpha}(y_i \mid x_i, z_i \in [\tilde{a}, \tilde{b}]) \right).$$
Then, we define the modified CML (MCML) estimator $\hat{\theta}_{MCML} = (\hat{\beta}_{MCML}, \hat{\alpha}_{MCML})$ by
\[
\hat{\theta}_{MCML} = \arg \max_{\theta} L_{MCML}(\theta).
\]
From (9) and Theorem 1, it is easy to show that
\[
n^{1/2} \left( W_{\tilde{a},\tilde{b}}(x, \beta, \alpha) - 1 \right) = O_p(1)
\]
and therefore
\[
n^{1/2} (p_{\beta,\alpha}(y \mid x, Z \in [\tilde{a}, \tilde{b}]) - f_{h(x)_{\beta,\alpha}}(y)) = O_p(1).
\]
Then, according to (11), both $L_{CML}(\theta)$ and $L_{MCML}(\theta)$ tend, under the model, to the unconditional likelihood function with rate $n^{-1/2}$. For this reason we conjecture that both the CML and the MCML estimator have the same asymptotic distribution than the unconditional ML estimator, that is,
\[
n^{1/2} (\hat{\theta}_{CML} - \theta_0) \to^D N_p(0, \mathcal{I}^{-1}(\theta_0)),
\]
and
\[
n^{1/2} (\hat{\theta}_{MCML} - \theta_0) \to^D N_p(0, \mathcal{I}^{-1}(\theta_0)),
\]
where $\to^D$ denotes convergence in distribution, $N_p(\mu, \Sigma)$ the $p$-variate normal distribution with mean $\mu$ and covariance matrix $\Sigma$, and $\mathcal{I}(\theta)$ the information matrix. This implies that $\hat{\theta}_{CML}$ and $\hat{\theta}_{MCML}$ are both fully efficient.

**Remark 1.** Empirical results show that, in order to optimize the finite sample efficiency, with no loss of robustness, it is convenient to iterate the conditional ML estimator as follows. Given a current value of $\hat{\theta}_{CML}$ (or $\hat{\theta}_{MCML}$), we compute new RQR-s. Then, we compute new values of $\tilde{a}$ and $\tilde{b}$ and use them to update $\hat{\theta}_{CML}$. Often, the process converges after a few iterations, but can also move away from the initial value. In the experiments reported in Section 4.1, we found that two steps are enough: the efficiency did not improve using more iterations. Moreover, in the
discrete case, the final estimator slightly depends on the sample \(\{u_1, \ldots, u_n\}\) used to compute \(\tilde{a}\) and \(\tilde{b}\). To remove this dependency, we propose to average the final step (MCML) over a few replications of this sample.

**Remark 2.** In certain circumstances, we may use a very simple alternative procedure to compute robust and consistent estimators of \(\beta_{01}, \alpha_0,\) and \(\eta_0\) in (5). We first identify a simple model, which is free of the dispersion parameter, and that can be taken as an approximation of (5). For example, the Poisson regression model with mean \(h(\beta_{01} + \eta_0\nu)\) may be taken as an approximation of the NB model. We then use an available robust procedure to estimate \(\beta_{01}\) and \(\eta_0\). In the NB case, the conditionally unbiased bounded influence estimators of Künsch et al. (1989), implemented in the R package “robeth” (Marazzi, 1992) is a natural choice. In the Beta regression case we note that Atkinson (1985) transforms the response so that the transformed dependent variable (e.g., \(\log(y/(1 - y))\)) assumes values on the real line, and then uses it in a linear regression analysis. Clearly, we may also use a robust regression estimator in this case, e.g., the MM estimator implemented in the R package “robustbase”.

Finally, we estimate \(\alpha_0\) using (7). Since the approximate model is not the correct one, the estimators do not converge to \(\beta_{01}, \alpha_0,\) and \(\eta_0\). Usual robust estimators converge however to their population values and can be used to define fixed cut-off values \(a\) and \(b\) for \(Z\), which also converge to their asymptotic values. The CML (MCML) estimator of \((\beta_{01}, \alpha_0, \eta_0)\) given \(Z \in [a, b]\) is then consistent under (5).

### 4 Simulation experiments for NB regression

We present simulation results only for the NB regression model (3)-(1). We compared the initial estimators \(\tilde{\beta}\) and \(\tilde{\alpha}\) and the final modified CML estimators \(\hat{\beta}\) and \(\hat{\alpha}\). In the following, these estimators will be referred as INI and CML respectively. All cutoff values were adaptive with \(\zeta_1 = 0.05\) and \(\zeta_2 = 0.95\). In order to compute the MRC
estimators we used the function maxCorGrid of the R package ccaPP (Alfons et al., 2015). The INI estimator was completed with the help of the weighted MT estimator described in Appendix 8. In order to estimate $\alpha_0$, we used the function $\psi$ defined by the equation for $\alpha$ of the optimal M estimator M80 described in Marazzi and Yohai (2010, p.174) and available in the “robustGLM” package. To compute the CML estimator, we used the standard R optimizer “optim”, reparametrizing $\alpha$ with $\sigma = \sqrt{\alpha}$ in order to satisfy the constraint $\alpha > 0$. (For a very small number of contaminated cases, the optimization process diverged; the initial solution was recorded in such cases.) Only two iterations of the CML procedure were computed. For comparison, we also computed the GM estimators of Aeberhard et al. (2014) that will referred as ACH in the following. To compute the ACH estimator, we used the R function glmrob.nb (available on internet) with the parameters: bounding.func='T/T', c.tukey.beta=4, c.tukey.sig=4, as suggested by the authors, and the option x-weight=hard that provides hard rejection weights for the covariate observations. We used the following model:

$$y \sim F_{\exp(x^T \beta_0, \alpha_0)}, \quad x = \begin{pmatrix} 1 \\ x^* \end{pmatrix}, \quad x^* \sim N(0, I_5), \quad (12)$$

$$\beta_0 = (1.5, 0.5, 0.25, 0, 0, 0), \quad \alpha_0 = 0.8.$$

### 4.1 Simulations at the nominal model

We first performed four experiments with samples of size $n = 100, 400, 1000, 2000$ from (12) without addition of outliers. For each experiment, the number of replications was $N = 1000$. To measure the quality of an estimator $(\beta, \alpha)$ we used the mean absolute estimation error (MAEE) and the mean absolute prediction error (MAPE). The MAEE of $\beta$ is defined by

$$\text{MAEE}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \|\beta_i^\# - \beta_0\|_1,$$
where $\beta_i^\#$ is the estimate of $\beta_0$ based on the $i^{th}$ replication and $||.||_1$ denotes the $l_1$ norm. The MAEE of $\alpha$ is defined in a similar way by

$$\text{MAEE}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} |\alpha_i^\# - \alpha_0|.$$ 

The MAPE of the prediction estimator $\mu_x = \exp(x^T \beta)$ of $\mu_{0,x} = \exp(x^T \beta_0)$ is defined as

$$\text{MAPE}(\mu) = \frac{1}{N} \sum_{i=1}^{N} |\mu_i^\# - \mu_{0i}|,$$

where $\mu_{0i} = \exp(x_{i}^{\#T} \beta_0)$ and $\mu_i^\# = \exp(x_{i}^{\#T} \beta_i^\#)$ and $x_i^\#$ is the $i^{th}$ replication of $x$.

Table 1 reports the empirical relative efficiencies measured as the ratios of the MAEE and MAPE of the robust estimators with respect to the corresponding MAEE and MAPE of the ML estimators.

| Table 1. Empirical relative efficiencies of coefficients, dispersion, and prediction estimators |
|---------------------------------------------------------------|
| We observe that the relative efficiencies of the initial estimators were low but could be improved with the help of the final MCML procedure. With the exception of the dispersion estimator for $n = 100$, our final estimator is much more efficient than the ACH competitor. (The tuning constants of the ACH estimator were apparently chosen by the authors in order to obtain a satisfactory degree of robustness.) |

4.2 Simulation with contaminated data

In another simulation the model (12) has been contaminated with 10% of pointwise contamination. Preliminary experiments showed that the estimators were quite sensitive to outlying values of $y$ when $x^* = (3, 1, 0, 0, 0)^T$. This value of $x$ is moderately outlying with respect to the majority of the covariate observations. Therefore, we
used point contaminations of the form \((x_{out}^*, y_{out})\) with \(x_{out}^* = (3, 1, 0, 0, 0)^T\) and a response \(y_{out}\) varying in the set \(\{0, 1, 2, 10, 20, 30, 40, 50, 60, 70, 100, 120, 180\}\). For each value of \(y_{out}\), we generated 1000 samples of size \(n = 400\) according to (12) and then replaced 10% of the observations with identical outliers of the form \((x_{out}^*, y_{out})\).

Table 2 reports the MAEE and MAPE of the estimators for the different values of \(y_{out}\). (Outliers were excluded in the computation of the MAPE). The results are also displayed in Figure 1. Both the MAPE and MAE of the proposed estimators were smaller than those of ACH for most values of \(y_{out}\).

| Table 2. MAEE and MAPE of coefficient, dispersion, and prediction estimators for varying \(y_{out}\). |

| Figure 1. Mean absolute prediction and estimation errors for varying \(y_{out}\). |

5 Application to hospital length of stay

In modern hospital management, stays are classified into “diagnosis related groups” (DRGs; Fetter et al., 1980) which are designed to be as homogeneous as possible with respect to diagnosis, treatment, and resource consumption. The mean cost of stay of each DRGs is periodically estimated with the help of administrative data on a national basis and used to determine “standard prices” for hospital funding and reimbursement. Typical stays are reimbursed according to the standard prices, whereas the reimbursement of exceptional stays (outliers) is subject to special negotiations among the partners. Since it is difficult to measure cost, length of stay (LOS) is often used as a proxy. Outliers are usually defined as observations with a LOS larger that some arbitrary cutoff value. In designing and refining the groups, the relationship between LOS and other variables which are usually available on administrative files has to be assessed and taken into account.
We first reconsider the example described in Marazzi and Yohai (2010). In this example there are not covariables, that is, only the parameters of a NB distribution are estimated. Table 3 shows the LOS of 32 stays classified into DRG “disorders of the nervous system” and we immediately identify three extreme values: 115, 198, 374 days. The arithmetic means with and without these observations are 25.5 and 4.4 days, respectively. We modeled the observed frequencies of LOS−1 (note that, by definition, the minimal LOS is 1) with a NB model. First, we computed the ML and the “optimal” M estimator referred as M80 in Marazzi and Yohai (2010). Then, we computed the modified CML estimator (called CML in the following) with $a = 0.05$ and $b = 0.95$ based on two iterations starting from M80, and averaged over 100 replications of $\{u_1,...u_n\}$. We also computed the three estimators (MLE*, M80*, CML*) after removal of the three outliers. The numerical results are shown in Table 4. They show that M80 and CML provided results which were similar to MLE* and unaffected by the outliers. The average values of $\tilde{a}$ and $\tilde{b}$ were $\tilde{a} = 0.044$ and $\tilde{b} = 0.953$ from which we derived $T_{\tilde{a}} = 1$, $T_{\tilde{b}} = 7$, $t_{\tilde{a}} = 0.61$, and $t_{\tilde{b}} = 0.43$. This means that, in the average, the CML estimator completely rejected LOS−1 values outside the interval [0, 8] and gives weights 0.61 and 0.57 to the extremes of this interval.

Table 3. Length of stay of 32 hospital patients.

Table 4. Estimates of LOS-1 mean and LOS-1 dispersion for disorders of the nervous system.
In a second example, we considered a sample of 649 hospital stays (256 male and 393 female patients) for the “major diagnostic category” (MDC) “Diseases and Disorders of the Endocrine, Nutritional And Metabolic System”. A MDC is simply a group of DRGs associated with a particular medical specialty. The data are shown in Figure 2 (two outliers with LOS = 84 and LOS = 122 fall beyond the upper limit of the figure).

We studied the relationship between LOS−1 and two covariates: Age of the patient \((x_1\) in years) and Sex of the patient \((x_2 = 0\) for males and \(x_2 = 1\) for females). We considered a NB model with exponential link and linear predictor \(\beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_3 x_1 x_3\). We compared the ML, the ACH, and the complete estimator (called CML in the following) proposed in Section 3. The ACH estimator was computed with the help of the R function glmrob.nb with the tuning parameters suggested by the authors. The CML step - with \(a = 0.05\) and \(b = 0.95\) and two iterations - was replicated 30 times with different vectors \(\{u_1, \ldots, u_n\}\). The average values of \(\tilde{a}\) and \(\tilde{b}\) were \(\bar{a} = 0.004\) and \(\bar{b} = 0.973\), from which we derived \(T_{\tilde{a},x_i}\), \(T_{\tilde{b},x_i}\), \(t_{\tilde{a},x_i}\), and \(t_{\tilde{b},x_i}\) (\(i = 1, \ldots, n\)). We found that 65 observations were totally rejected, 62 fell on the lower limits \(T_{\tilde{a},x_i}\) (receiving an average weight 0.96) and 9 on the upper limits \(T_{\tilde{b},x_i}\) (with an average weight 0.52). In Figure 2, the full outliers are marked by cross signs (x) and the borderline observations by plus signs (+). Thus, we had about 11% of contamination, mostly located on the upper tail of the LOS distribution; no leverage point in the covariate space were present in these data. We also computed the ML estimator (ML*) after removal of the full outliers. The numerical results are given in Table 5 and the prediction lines are drawn in Figure 2.

*Figure 2. Data: LOS-1 and Age of 649 patients and fitted models according to CML and ML.*

*Table 5. Coefficient (standard errors) and dispersion estimates for disorders of the endocrine system.*
We observe that the CML and the ML* coefficient estimates are very close and quite similar to the ACH estimates. (However, the standard errors provided by glmrob.nb are surprisingly large.) We also note that the dispersion parameter is heavily inflated by the contamination. For CML and ML*, the Sex effect ($\beta_2$) is significant at the 5% level and the interaction ($\beta_3$) is not significant. Instead, for ML, the interaction is significant at the 5% level, but not the effect of Sex. Thus, the classical and the robust inferences are different.

Figure 3 shows three uniform qq-plots of randomized tail probabilities $z_1, \ldots, z_n$ based on different estimates of $\alpha$ and $\beta$. In panel (a) the ML estimator has been used and the sigmoidal shape suggests that the estimated model is incorrect. In panel (b), the $z$-values were based on the modified CML estimator; the plot is more linear but it gradually departs form the diagonal for increasing quantiles. This suggests that the robustly fitted model is adequate for a large proportion of data but not for those corresponding to very large values of $z$. Panel (c) is based on ML* and the $z$-values corresponding to the full outliers based on CML have been removed from the plot; this plot follows the diagonal line very well. Finally, the boxplots in panel (d) compare the distribution of the absolute residuals (without outliers) based on ML, ACH, CML, and ML*; the two latter ones are globally smaller than the former ones. We conclude that CML (and ML*) provide an adequate model for about 90% of the population.

*Figure 3. qq-plots of randomized tail probabilities based on ML, CML, ML with removal of the extreme z-values from the plot, and boxplots of the absolute residuals of ML, ACH, CML, and ML*.*
6 Discussion

In many areas of applied statistics, the data may be affected by a high level of contamination. An example is the analysis of hospital length of stay, where contamination levels as high as 10% are not uncommon. For this reason, different ad hoc rules of trimming had long been used by practitioners to remove outliers (e.g., Marazzi et al., 1998) from their data. In these applications, well founded highly robust procedures are needed.

Maronna et al. (1979) showed that classical M and GM estimators of regression (see e.g., Huber, 1980, Hampel et al., 1986) were unable to combine a high level of robustness and a high level efficiency: M and GM estimators can be very efficient, but are very sensitive to outliers in the factor space. This work stimulated the research on high breakdown-point estimation that provided LMS, LTS, and S estimators (see e.g., Maronna et al., 2006) just to mention three among many other procedures. Then, for the usual linear regression problem, the MM estimators of Yohai (1987) combined high breakdown point and high efficiency with the help a two step approach: in the first step, a very robust initial fit (an S estimator) provided the tool for outlier identification; the second step was based on an efficient estimator (an M estimator), where the outliers were downweighted. Since then, similar two-step procedures have been proposed for different models (Marazzi and Yohai, 2004; Locatelli et al. 2010; Agostinelli et al., 2014).

However, the familiar high breakdown point regression estimators used in the first step are based on minimization of a robust measure of the residual scale and, unfortunately, cannot be used for GLMs with a dispersion parameter, such as NB and Beta regression. The reason is that the residual distribution depends on the covariates and robust residual measures of scale are not available in this case. In this paper, we propose a more general approach that bypasses residual scales.

Our proposal is an original assembly of well known procedures. In the initial step
we use the MRC estimator (Han, 1987) to estimate the slopes up to a scale factor. A very fast algorithm to compute this estimator has recently been proposed in Alfons et al (2016). We complete the MRC with the help of a weighted MT estimator (Valdora and Yohai, 2014) of a simple negative binomial regression. We then use randomized quantile residuals (Dunn and Smyth, 1996) to determine adaptive cutoff values $\tilde{a}$ and $\tilde{b}$ using a procedure similar to the one proposed in Marazzi and Yohai (2004). Influential outliers are identified by the residuals not belonging to $[\tilde{a}, \tilde{b}]$. Finally, we compute a conditional ML estimator where residuals belong to $[\tilde{a}, \tilde{b}]$. Since, in the absence of outliers, $\tilde{a} \to 0$ and $\tilde{b} \to 1$, the CML estimator tends to the ML estimator for $n \to \infty$. It is therefore fully efficient.

Monte Carlo simulations confirm that our proposal is very efficient under the model and very robust under point contamination, both in the response and the covariate distributions. This kind of contamination is unrealistic; however, it is generally the least favorable one and allows evaluation of the maximal bias an estimator can incur. The CML estimator for NB regression also resists to a moderate fraction of excess zeroes in the response. A more vigorous treatment of this peculiarity of count data should however be approached with the help of specific models, such as hurdle models (see, e.g. Min and Agresti, 2002, and Cantoni and Zedini, 2009).

We have shown that the proposed method is a useful tool for modelling hospital length of stay as a function of available covariates, while identifying influential outliers according to a model based rule. A set of R functions to compute the proposed estimators is made available as an R package.
Appendices

7 Proof of (9)

To simplify notations, we just consider the case without covariates; the extension to the regression case is straightforward. We suppose that $\vartheta = (\mu, \alpha)$ is given and let $z = F_\vartheta(y) - u f_\vartheta(y)$, where $u \sim U[0, 1]$. Suppose that $a$ and $b$ are given cutoff values for $z$ and define, for any $c$,

$$y^*(c) = \max\{y : F_\vartheta(y) \leq c\}.$$

Note that

$$F_\vartheta(y^*(a) + 1) - u f_\vartheta(y^*(a) + 1) \geq a$$

is equivalent to

$$u \leq \frac{F_\vartheta(y^*(a) + 1) - a}{f_\vartheta(y^*(a) + 1)} = t_a.$$

Similarly

$$F_\vartheta(y^*(b) + 1) - u f_\vartheta(y^*(b) + 1) \leq b$$

is equivalent to

$$u \geq \frac{F_\vartheta(y^*(b) + 1) - b}{f_\vartheta(y^*(b) + 1)} = t_b.$$

Put $T_a = y^*(a) + 2$, $T_b = y^*(b)$, and $A = \{y : T_a \leq y \leq T_b\}$. We have

$$\{a \leq z \leq b\} = A \cup \{y = T_a - 1, \ u \leq t_a\} \cup \{y = T_b + 1, \ u \geq t_b\},$$

and then

$$P_\vartheta(a \leq z \leq b \mid u) = P_\vartheta(A) + f_\vartheta(T_a - 1)I(u \leq t_a) + f_\vartheta(T_b + 1)I(u \geq t_b),$$
where $P_\vartheta(A) = F_\vartheta(T_b) - F_\vartheta(T_a - 1)$. Let $v = I(a \leq z \leq b)$. Since $E[I(u \leq t_a)] = P(u \leq t_a) = t_a$, the distribution of $y \mid v = 1$ is given by

$$p_\vartheta(y \mid v = 1) = \begin{cases} \frac{f_\vartheta(y)}{Q(\vartheta)} & \text{if } y \in A, \\ \frac{f_\vartheta(y) t_a}{Q(\vartheta)} & \text{if } y = T_a - 1, \\ \frac{f_\vartheta(y)(1 - t_b)}{Q(\vartheta)} & \text{if } y = T_b + 1, \\ 0 & \text{if elsewhere.} \end{cases}$$

where

$$Q(\vartheta) = E[P_\vartheta(a \leq z \leq b \mid u)] = P_\vartheta(A) + p_\vartheta(T_a - 1)t_a + p_\vartheta(T_b + 1)(1 - t_b).$$

### 8 Weighted MT estimator of simple regression

We describe the use of the weighted MT estimator to compute $\tilde{\beta}_1^*(\alpha)$ and $\tilde{\eta}^*(\alpha)$ introduced in subsection 3.1. We consider the simple regression model $Y \mid v \sim F_{h(\beta_0 + \eta_0v), \alpha}$. Assuming that $\alpha$ is known, the weighted MT estimator of $(\beta_0, \eta_0)$ is defined as follows.

$$(\tilde{\beta}_1^*(\alpha), \tilde{\eta}^*(\alpha)) = \arg \min_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^{n} w(x_i, \hat{M}, \hat{S}) \rho(t(y_i, \alpha) - m(h(\beta_0 + \beta_1 x_i, \alpha)),$$  \quad (13)

where $\rho$ is a continuous and bounded function with a unique local minimum at 0, $m$ is the function defined by

$$m(\mu, \alpha) = \arg \min_{\gamma} E_{\mu, \alpha} (\rho(t(y, \alpha) - \gamma)),$$  \quad (14)

t(y, \alpha)$ is a variance stabilizing transformation and $w(x, \hat{M}, \hat{S})$ is a nonnegative non-increasing function of $|x - \hat{M}|/\hat{S}$, where $\hat{M}$ and $\hat{S}$ are robust estimators of location.
and scale of the covariate $x$. Usually, $\rho$ is taken in the Tukey’s biweight family given by

$$
\rho_c^T(u) = 1 - \max \left( 1 - \left( \frac{u}{c} \right)^2 , 1 \right)^3.
$$

In our simulations in Section 4 and the examples in section 5 with the NB distribution, we used the transformation

$$
t(y, \alpha) = \begin{cases} 
  s(y, \alpha) & \text{if } 0 < \alpha < 1.3 \\
  s(y, 1.3) & \text{if } \alpha > 1.3
\end{cases},
$$

where

$$
s(y, \alpha) = \sqrt{\frac{1}{\alpha} - 0.5 \arcsinh \left( \sqrt{\frac{y + 3/8}{1/\alpha - 3/4}} \right)}.
$$

This is a modification of the transformation proposed by Yu (2009) to allow values of $\alpha$ larger than $4/3$. We take $w(x, \hat{M}, \hat{S}) = I(|x - \text{median}(x_i)|/\text{mad}(x_i) < 2)$ and $h(z) = \exp(z)$. Since the variance of $t(y, \alpha)$ is almost constant, it is not necessary to divide the argument of $\rho_c$ by a scale estimator. While the efficiency of the estimator increases with $c$, its degree of robustness decreases. Since the weighted MT estimator is used to define an initial estimator whose efficiency will be improved in further steps, the value of $c$ is chosen in order to obtain a satisfactory degree of robustness. By trial and error we obtain the following rule for choosing $c$ as a function of $\alpha$: $c = 1.5 \sigma(\alpha)$, where, for each $\alpha$, $\sigma(\alpha)$ is the constant that approximates the standard deviation of $t(y, \alpha)$. The value of $\sigma(\alpha)$ is obtained by interpolation the values in the following Table A1:

| $\alpha$ | .10 | .20 | .30 | .40 | .50 | .60 | .70 | .80 | .90 | 1.0 | 1.1 | 1.2 | 1.3 |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $\sigma$ | .41 | .40 | .39 | .37 | .36 | .35 | .33 | .32 | .30 | .29 | .27 | .26 | .24 |

Table A1. Approximated standard deviations of $t(y, \alpha)$ for the NB distribution
For the Beta distribution, we have $\text{Var}_{\mu,\alpha}(y) = \mu(1 - \mu)/(1 + \alpha)$ and a suitable variance stabilizing transformation (Bartlett, 1947) is given by

$$t(y, \alpha) = \int_0^y \frac{1}{\text{Var}_{\mu,\alpha}(y)^{1/2}} d\mu = \sqrt{1 + \alpha \arcsin(\sqrt{y})}.$$ 

In our experiments we used this transformation for $\alpha \in [5, 50]$ and link function $h(u) = \exp(u)/(1 + \exp(u))$. We follow the same approach as in the NB case. The values of the approximated variances can be found in the following Table A2:

| $\alpha$ | 5  | 6  | 7  | 8  | 9  | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |
|----------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| $\sigma$ | .42 | .43 | .43 | .44 | .45 | .45 | .47 | .48 | .48 | .49 | .49 | .49 | .49 | .49 |

Table A2. Approximated standard deviations of $t(y, \alpha)$ for the beta distribution

When $\alpha$ is unknown, the estimator $(\tilde{\beta}_1^*(\tilde{\alpha}), \tilde{\eta}^*(\tilde{\alpha}))$ simultaneously satisfies equations (13) and (7). To compute an approximate solution we consider a grid of possible values of $\alpha$, namely the values in the tables above. For each $\alpha$ in the grid, we first compute $(\tilde{\beta}_1^*(\alpha), \tilde{\eta}^*(\alpha))$ and then the solution $\tilde{\alpha}^*$ of (7). The desired approximation is then defined as the vector $(\tilde{\beta}_1^*(\tilde{\alpha}^*), \tilde{\eta}^*(\tilde{\alpha}^*))$ for which the difference between $\alpha$ and $\tilde{\alpha}^*$ is minimal.

9 Proof of Theorem 1

We consider the discrete case, where the RQRs are defined by

$$z_i = F_{h(x^T\tilde{\beta}, \tilde{\alpha}^*)}(y_i) - u_i f_{h(x^T\tilde{\beta}, \tilde{\alpha}^*)}(y_i), \ 1 \leq i \leq n,$$

By Assumption B, there exist $A_0$ and $B_0$ such that, if

$$D_n = \{n^{1/2}|\tilde{\alpha} - \alpha_0| \leq A_0, \ ||n^{1/2}(\tilde{\beta}^T - \beta_0)|| \leq B_0\},$$

26
we have
\[ P(D_n) \geq 1 - \zeta_1/2. \]

Put
\[ v_i = F_{h(x^T \beta_0), \alpha_0} - u_i f_{h(x^T \beta_0), \alpha_0}(y_i), \quad 1 \leq i \leq n. \]

Then the \( v_i \)'s are i.i.d. with distribution \( U[0, 1] \). By Assumption C, there exist \( K_1 \) and \( K_2 \) such that
\[ z_i \geq v_i - \left( K_1 |\tilde{\beta} - \beta_0| + K_2 |\tilde{\alpha} - \alpha_0| \right) \]
i.e.,
\[ z_i \geq v_i - n^{-1/2} B_n, \]
where \( B_n = O_P(1) \). Let \( e_0 \) such that, if \( M_n = \{ B_n \leq e_0 \} \), then
\[ P(M_n) \geq 1 - \zeta_1/2. \]

Let \( F_{zn} \) and \( F_{vn} \) be the empirical distributions of the \( z_i \)'s and \( v_i \)'s respectively. Then, we have
\[ F_{zn}(v) \leq F_{vn}(v + n^{-1/2} B_n). \quad (15) \]

Since
\[ E_n = \sup_v n^{1/2} |F_{vn}(v) - v| = O_P(1), \]
we get \( F_{vn}(v) \leq v + n^{-1/2} E_n \). Then, putting \( G_n = B_n + E_n \), by (15) we obtain
\[ F_{zn}(v) \leq v + n^{-1/2} G_n. \quad (16) \]

In a similar way we get
\[ F_{zn}(v) \geq v - n^{-1/2} G^*_n. \quad (17) \]
where \( G^*_n = O_p(1) \). Let
\[ H_{na}(v) = \frac{\sup(F_{zn}(v) - F_{zn}(a), 0)}{1 - F_{zn}(a)} \]
and
\[ A = \{ a : \sup_{v \leq \zeta_1} (H_{na}(v) - v) \leq 0 \} . \]

Then
\[ \tilde{a} = \inf A \]

Note that \( a \in A \) is equivalent to

\[ F_{zn}(v) \leq v(1 - F_{zn}(a)) + F_{zn}(a) \text{ for all } v \leq \zeta_1 \]

and this is equivalent to

\[ F_{zn}(a)(1 - v) \geq F_{zn}(v) - v \text{ for all } v \leq \zeta_1 . \]

By (16) and (17) a sufficient condition for \( a \in A \) is that

\[ (a - n^{-1/2}G^*_n)(1 - \zeta_1) \geq n^{-1/2}G_n \]

or equivalently that

\[ a \geq n^{-1/2} \left( \frac{G_n}{1 - \zeta_1} + G^*_n \right) \]

This implies that

\[ \tilde{a} \leq n^{-1/2} \left( \frac{G_n}{1 - \zeta_1} + G^*_n \right) . \]

proving that \( n^{1/2} \tilde{a} \) is bounded in probability. The proof that \( n^{1/2}(\tilde{b} - 1) \) is bounded in probability too is similar.
References

Abrevaya J. (1999). Computation of the maximum rank correlation estimator. Economics Letters, 62, 279–285.

Aeberhard W.H., Cantoni E. and Heritier S. (2014). Robust inference in the negative binomial regression model with an application to falls data. Biometrics, 70, 920-931. DOI: 10.1111/biom.12212

Agostinelli C., Marazzi A. and Yohai V.J. (2014). Robust estimators of the generalized log-gamma distribution. Technometrics, 56(1), 92-101.

Alfons A., Croux C. and Filzmoser P. (2016). Robust maximum association estimators. Journal of the American Statistical Association. In press.

Alfons A. (2015). ccaPP: (Robust) canonical correlation analysis via projection pursuit. R package version 0.3.1, URL http://CRAN.R-project.org/package=ccaPP.

Amiguet, M. (2011). Adaptively weighted maximum likelihood estimation of discrete distributions. Ph.D. thesis, Université de Lausanne, Switzerland.

Austin, P.C., Rothwell, D.M. and Tu, J.V. (2002). A comparison of statistical modeling strategies for analyzing length of stay after CABG surgery. Health services & outcomes research methodology (3), 107-133. DOI:10.1023/A:1024260023851

Cadigan N. G. and Chen J. (2001). Properties of robust M–estimators for Poisson and negative binomial data. Journal of Statistical Computation and Simulation, 70, 273-288.

Cantoni E., and Ronchetti E. (2001). Robust inference for generalized linear models. Journal of the American Statistical Association, 96(455),1022-1030.

Cantoni E. and Zedini A. (2009). A robust version of the hurdle model. Cahiers du département d’économétrie No 2009.07, Faculté des sciences économiques et sociales, Université de Genève.
Carter E.M. and Potts H.W.W. (2014). Predicting length of stay from an electronic patient record system: a primary total knee replacement example. BMC Medical Informatics & Decision Making 14: 26. DOI:10.1186/1472-6947-14-26.

Cribari–Neto F. and Zeiles, A. (2010). Beta regression in R. Journal of Statistical Software, 34, 1–24.

Cuesta-Albertos J.A., Matrán C. and Mayo-Iscar A (2008). Trimming and likelihood: robust location and dispersion estimate in the elliptical model. The Annals of Statistics, 36(5), 2284–2318.

Davison A.C. and Snell E.J. (1991). Residuals and diagnostics. In Statistical Theory and Modelling: In Honour of Sir David Cox. D.V. Hinkley, N. Reid and E.J. Snell (editors), 83–106. Chapman and Hall.

Dunn P.K. and Smyth G.K. (1996). Randomized quantile residuals. Journal of Computational and Graphical Statistics, 5(3), 236-244.

Espinheira P. L., Ferrari S. L. P. and Cribari-Neto, F. (2008). Influence diagnostics in beta regression. Computational Statistics & Data Analysis, 52(9), 4417-4431.

Ferrari S. L. P. and Cribari-Neto F. (2004). Beta regression for modelling rates and proportions. Journal of Applied Statistics, 31(7), 799–815

Fetter R.B., Shin Y., Freeman J.L., Averill R.F., and Thompson J.D. (1980). Casemix definition by diagnosis-related groups. Medical care, 18(1), 1-53.

Gervini D. and Yohai V.J. (2002). A class of robust and fully efficient regression estimators. Annals of Statistics, 30(2), 583-616.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J. and Stahel W.A. (1986). Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Han A.K. (1987a). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. Journal of Econometrics, 35(23), 303-316.
Han A.K. (1987b). A non-parametric analysis of transformations. Journal of Econometrics, 35, (2-3), 191-209.

Hilbe J.M. (2008). Negative Binomial Regression. Cambridge University press.

Huber P.J. (1980). Robust Statistics. Wiley, New York.

Hunger M., Baumert J. and Holle R. (2011). Analysis of SF-60 index data: is beta regression appropriate? Value in Health 14, 759-767.

Künsch H.R., Stefanski L.A. and Carroll R.J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. Journal of the American Statistical Association, 84(406), 460-466.

Locatelli I., Marazzi A. and Yohai V.J. (2010). Robust accelerated failure time regression. Computational Statistics & Data Analysis, 55(1), 874-887.

Marazzi A. (1993). Algorithms, Routines, and S-Functions for Robust Statistics. Wadsworth, Inc., Belmont, California.

Marazzi A., Paccaud F., Ruffieux C. and Beguin C. (1998). Fitting the distribution of length of stay by parametric models. Medical Care, 36(6), 915-927.

Marazzi A. and Yohai V.J. (2004). Adaptively truncated maximum likelihood regression with asymmetric errors. Journal of Statistical Planning and Inference. 122 (1-2), 271-291.

Marazzi A. and Yohai V.J. (2010). Optimal robust estimates based on the Hellinger distance. Advances in Data Analysis and Classification. Springer-Verlag.

Maronna R.A., Martin R.D. and Yohai V.J. (2006). Robust Statistics Theory and Methods. Wiley & Sons, Ltd.

Maronna R., Bustos O. and Yohai V.J. (1979). Bias-and efficiency-robustness of general M-estimators for regression with random carriers, Smoothing techniques for curve estimation, 91-116.
Min Y. and Agresti A. (2002). Modeling nonnegative data with clumping at zero: A survey. Journal of the Iranian Statistical Society, 1,(1-2), 7-33.

Nelder J.A. and Wedderburn R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society, Series A, 135 (3), 370-384.

Venables W. N. and Ripley B. D. (1999). Modern Applied Statistics with S-PLUS. Third edition. Springer.

Rocha, A. V. and Simas, A. B. (2010). Influence diagnostics in a general class of beta regression models. Test, 20(1), 95-119.

Seow W.J., Pesatori A.C., Dimont E., Farmer P.B., Albetti B, et al. (2012). Urinary benzene biomarkers and DNA methylation in Bulgarian petrochemi workers: Study findings and comparison of linear and beta regression models. PLOS ONE 7: e50471.

Sherman R.P. (1993). The limiting distribution of the maximum rank correlation estimator. Econometrica, 61(1), 123-137.

Swearingen C.J., Tillley B.C., Adams R.J., Rumboldt Z., Nicholas J.S., Bandyopadhyay D. and Woolson R.F. (2011). Application of Beta Regression to Analyze Ischemic Stroke Volume in NINDS rt-PA Clinical Trials. Neuroepidemiology, 37(2), 73-82.

Valdora M. and Yohai V.J. (2014). Robust estimation in generalized linear models. Journal of Statistical Planning and Inference, 146, 31-48.

Yu G. (2009). Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. Statistics and Probability Letters, 79, 1621-1629.

Yohai V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. Annals of Statistics, 15(2), 642-656.
Table 1. Empirical relative efficiencies of coefficients, dispersion, and prediction estimates.

| $n$  | $\beta$ | $\alpha$ | $\mu$ |
|------|---------|----------|-------|
|      | INI     | CML      | ACH   | INI     | CML      | ACH   | INI     | CML      | ACH   |
| 100  | 0.55    | 0.74     | 0.71  | 0.78    | 0.71     | 0.76  | 0.50    | 0.70     | 0.76  |
| 400  | 0.52    | 0.88     | 0.75  | 0.73    | 0.85     | 0.79  | 0.48    | 0.89     | 0.76  |
| 1000 | 0.51    | 0.93     | 0.78  | 0.73    | 0.93     | 0.83  | 0.48    | 0.93     | 0.78  |
| 2000 | 0.54    | 0.95     | 0.75  | 0.73    | 0.94     | 0.83  | 0.50    | 0.95     | 0.75  |

Table 2. MAEE and MAPE of coefficient, dispersion, and prediction estimates for varying $y_{out}$.

| $y_{out}$ | 0 | 1 | 2 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 100 | 120 | 180 |
|-----------|---|---|---|----|----|----|----|----|----|----|-----|-----|-----|
| $\beta$  | INI | 0.86 | 0.57 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 |
|          | CML | 0.72 | 0.76 | 0.78 | 0.55 | 0.38 | 0.32 | 0.33 | 0.37 | 0.42 | 0.45 | 0.53 | 0.55 | 0.48 |
|          | ACH | 1.27 | 1.19 | 1.09 | 0.67 | 0.46 | 0.41 | 0.45 | 0.50 | 0.55 | 0.59 | 0.70 | 0.75 | 0.88 |
| $\alpha$ | INI | 0.45 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 | 0.09 |
|          | CML | 0.45 | 0.25 | 0.13 | 0.18 | 0.23 | 0.23 | 0.23 | 0.21 | 0.20 | 0.19 | 0.15 | 0.12 | 0.09 |
|          | ACH | 0.53 | 0.12 | 0.10 | 0.27 | 0.29 | 0.29 | 0.27 | 0.26 | 0.24 | 0.23 | 0.19 | 0.17 | 0.12 |
| $\mu$    | INI | 1.92 | 1.17 | 1.11 | 1.11 | 1.11 | 1.11 | 1.11 | 1.11 | 0   | 1.11 | 1.11 | 1.11 | 1.11 |
|          | CML | 1.59 | 1.91 | 1.69 | 1.19 | 0.79 | 0.62 | 0.63 | 0.74 | 0.88 | 1.04 | 1.40 | 1.54 | 1.34 |
|          | ACH | 2.68 | 2.39 | 2.18 | 1.40 | 0.97 | 0.78 | 0.78 | 0.89 | 1.04 | 1.19 | 1.64 | 1.91 | 2.63 |

33
| LOS | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 16 | 115 | 198 | 374 |
|-----|---|---|---|---|---|---|---|---|---|----|-----|-----|-----|
| frequency | 2 | 6 | 5 | 5 | 4 | 2 | 1 | 1 | 1 | 1   | 1    | 1    | 1    |

Table 3. Length of stay of 32 hospital patients.

| MLE | M80 | CML | MLE* | M80* | CML* |
|-----|-----|-----|------|------|------|
| $\mu$ | 24.47 | 3.58 | 3.12 | 3.41 | 3.17 | 3.39 |
| $\alpha$ | 3.08 | 0.44 | 0.32 | 0.35 | 0.24 | 0.42 |

Table 4. Estimates of LOS-1 mean and LOS-1 dispersion for disorders of the nervous system.

| Beta Coefficients | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\alpha$ |
|-------------------|---------|---------|---------|---------|---------|
| ML                | 1.266   | 0.017   | 0.064   | -0.009  | 1.067   |
|                   | (0.134) | (0.002) | (0.178) | (0.003) | (0.067) |
| ACH               | 1.656   | 0.004   | -1.055  | 0.012   | 0.542   |
|                   | (0.726) | (0.011) | (0.735) | (0.011) | (—)     |
| CML               | 0.899   | 0.017   | -0.269  | -0.002  | 0.593   |
|                   | (0.113) | (0.002) | (0.154) | (0.003) | (0.049) |
| ML*               | 0.846   | 0.016   | -0.253  | -0.002  | 0.503   |
|                   | (0.114) | (0.002) | (0.156) | (0.003) | (0.046) |

Table 5. Coefficient (standard errors) and dispersion estimates for disorders of the endocrine system.
Figure 1. Mean absolute prediction and estimation errors for varying $y_{out}$. 
Figure 2. Data: LOS and Age of 649 patients. Black circles are men, gray circles are women. Full outliers are marked by cross signs (x); borderline observations by plus signs (+). Fitted models according to CML (solid lines) and ML (broken lines): black for men, gray for women.
Figure 3. qq-plots of randomized tail probabilities based on: ML (panel a), CML (panel b), ML with removal of the largest z-values from the plot (panel c). Panel (d): boxplots of the absolute residuals of ML, ACH, CML, and ML*.