Robust estimators for negative binomial regression.

M. Amiguet[1], A. Marazzi[1], M.Valdora[2] and **V.J. Yohai** [2*]

[1] *Université de Lausanne; Michael.Amiguet@chuv.ch, alfio.marazzi@chuv.ch*
[2] *Universidad de Buenos Aires; mvaldora@gmail.com, vyohai@dm.uba.ar*
[*] *Presenting author*

Keywords. *Kendall rank correlation coefficient; Full efficiency; Conditional maximum likelihood estimator*

In recent years, negative binomial (NB) regression has received increasing attention as a tool for modeling count data in presence of overdispersion. A convenient way to parametrize the negative binomial probability function is by the mean μ and the dispersion parameter α. We denote this distribution by NB(μ, α)

The negative binomial regression model assumes that we observe a response y and a vector of covariables $\mathbf{x} \in \mathbf{R}^p$, so that $y|\mathbf{x}$ has distribution NB($h(\beta_0^{\mathrm{T}}\mathbf{x} + \delta), \alpha_0)$), where the link function h is known while $\beta_0 = (\beta_{01}, ..., \beta_{0p})$ and α_0 are unknown parameters.

One way to estimate these parameters is by means of the maximum likelihood estimator. However these estimators are very sensitive to the presence of outliers in the sample. A robust estimator for this model was proposed by Aeberhard, Cantoni, and Heritier [2014].

We are going to introduce a new estimator which is simultaneously highly robust and fully efficient. Suppose tat we have a sample $(y_1, \mathbf{x}_1), ..., (y_n, \mathbf{x}_n)$, then the estimator we propose is defined by the following three steps.

First step. We first obtain a consistent initial estimate of $\beta_0^* = \beta_0/||\beta_0||_2$. This estimator estimator is defined by

$$\widetilde{\beta}^* = \arg\min_{\beta^*} \tau((y_1, ..., y_n)), (\beta^{*\mathrm{T}}\mathbf{x}_1), ..., (\beta^{*\mathrm{T}}\mathbf{x}_n)),$$

where if $\mathbf{z} = (z_1, ...z_n)$ and $\mathbf{w} = (w_1, ..., w_n)$, we call $\tau(\mathbf{z}, \mathbf{w})$ the Kendall rank correlation between $\mathbf{z}$ and $\mathbf{w}$ given by

$$\tau(\mathbf{z}, \mathbf{w}) = \#\left\{(i, j), 1 \le i < j \le n : \mathrm{sign}\left((z_i - w_i)(z_j - w_j)\right) \ge 0\right\}.$$

This estimator was proposed first by Han [1987] for another models. It may be proved that this estimator is consistent for any strictly increasing link function h.

Second step. In this step we obtain initial estimators of $\eta_0 = ||\beta_0^*||$, δ_0 and α_0. For that purpose, let $\widehat{z}_i = \widetilde{\beta}^{*\mathrm{T}}\mathbf{x}_i$, $1 \leq i \leq n$. Then we fit the following negative binomial regression model with just one covariable : the distribution of $y_i|\widehat{z}_i$ is $\mathrm{NB}(h(\eta_0\widehat{z}_j + \delta_0))$. Observe that this model holds exactly if we replace $\widehat{z}_i$ by $z_i = \beta_0^{*\mathrm{T}}\mathbf{x}_i$. Then we obtain estimators $\widetilde{\eta} = ||\gamma_0||$, $\widetilde{\delta}$ and $\widetilde{\alpha}_0$ using an M-estimator similar to the ones proposed by Marazzi and Yohai [2004] to estimate(μ_0, α_0) given a sample of a $\mathrm{NB}(\mu_0, \alpha_0)$ distribution. Finally, we complete the initial estimating by taking $\widetilde{\beta}{=}\widetilde{\alpha}\widetilde{\beta}^*$ as estimator of β_0

Third step. We transform the variables y_i as follows. Put $\theta = (\mu, \alpha)$ and let $p(., \theta)$ and $F(., \theta)$ the probability and distribution functions of the $\mathrm{NB}(\mu, \alpha)$ distribution respectively. Let y with $\mathrm{NB}(\mu, \alpha)$ distribution and $v = F(y, \theta) - up(y, \theta)$, where u has uniform distribution in $[0, 1]$ (U(0,1)) and is independent of y. Then v has U(0,1) distribution. Call $\widetilde{\theta}_i = (\widetilde{\beta}^{\mathrm{T}}\mathbf{x}_i + \widetilde{\delta}, \widetilde{\alpha})$ and let $u_1, ..., u_n$ be i.i.d. U(0,1) variables which are independent of the sample. Then $v_i = F(y_i, \widetilde{\theta}_i) - u_i p(y, \widetilde{\theta}_i)$, $1 \leq i \leq n$, are approximately i.i.d. U(0,1) variables. Then we can detect outliers comparing the empirical distribution of $r_i = |v_i - 0.5|, 1 \leq i \leq n$, with the distribution $F_0(u) = 2uI([0, 1])$ of $|u-0.5|$, where u has distribution U(0,1). Let $r_{(1)} <, ..., < r_{(n)}$ be the ordered sample and for $t = 1, ..., n$ let H_t be the empirical distribution of $r_{(1)}, ..., r_{(t)}$. Put $s_0 = \min\{s : \min_{r \geq 0.5-\varepsilon}(H_{n-s}(r) - H_0(r)) \geq 0\}$, where ε is a small number, e.g., $\varepsilon = 0.05$. Then, the observations such that $|v_i - 0.5| > r_{(n-s)} = \phi$ are going to be considered outliers and eliminated from the sample

Then, the final estimators are defined by

$$(\widehat{\beta}, \widehat{\delta}, \widehat{\alpha}) = \arg\min_{\beta,\alpha,\delta} L\left(y_1, y_2, ..., y_n, \beta, \alpha, \delta | \, \mathbf{x}_1, .., \mathbf{x}_n, \max_{1 \leq i \leq n} v_i \leq \phi\right),$$

where $L(y_1, ..., y_n, \beta, \alpha, \delta|\mathbf{t})$ denotes the conditional likelihood of $\mathbf{y} =(y_1, y_2, ..., y_n)$ given $\mathbf{t}$, when the parameters are β, α, δ. It can be proved that under the model we have $s_0/n \to 0$. We show that this implies that the final estimators are fully efficient. Moreover, a Monte Carlo simulation study show that they are also highly robust.

References

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.

Han, A.K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, **35**, 303-316.

Marazzi, A. and Yohai, V.J. (2010). Optimal robust estimates based on the Hellinger distance. *Advances in Data Analysis and Classification,* **4**, 169-179.