Small-sample likelihood inference in extreme-value regression models

Eliane C. Pinheiro1 2
Silvia L.P. Ferrari1 2

1 Introduction

The extreme-value distribution is frequently used to model extreme events, such as extreme floods and wind speed, and in survival or reliability analysis to model the logarithm of lifetime data. In this paper we deal with a general class of extreme-value regression models introduced by Barreto-Souza and Vasconcellos (2011). The authors considered the issue of correcting the bias of the maximum likelihood estimators in small samples. Here, our goal is to derive Skovgaard’s adjusted likelihood ratio statistics in this class of models. We show that the adjustment term has a simple compact form that can be easily implemented from standard statistical software. The adjusted statistic is approximately distributed as $\chi^2$ with a high degree of accuracy. Also, we aim to compare the finite sample performance of the three classical tests, the likelihood ratio, Wald and score tests, the gradient test, which has been recently proposed by Terrell (2002), and the adjusted likelihood ratio test obtained in this paper.

2 Model

Let $y_1, \ldots, y_n$ be independent random variables, where each $y_t$, $t = 1, \ldots, n$, has an extreme-value distribution with parameters $\mu_t$ and $\phi_t$ and density function

$$f(y; \mu_t, \phi_t) = \frac{1}{\phi_t} \exp \left( - \frac{y - \mu_t}{\phi_t} \right) \exp \left( - \exp \left( - \frac{y - \mu_t}{\phi_t} \right) \right), \quad y \in \mathbb{R},$$

(1)

where $\mu \in \mathbb{R}$ and $\phi > 0$ are the location and dispersion parameters, respectively. The mean and the variance of $y_t$ are

$$\mathbb{E}(y_t) = \mu_t + \mathcal{E}\phi_t$$

and

$$\text{var}(y_t) = \frac{\pi^2}{6} \phi_t^2,$$

1Departamento de Estatística, Universidade de São Paulo. e-mail: elianecp@ime.usp.br
2Agradecimento ao CNPq pelo apoio financeiro.
respectively, where $\mathcal{E}$ is the Euler constant; $\mathcal{E} \approx 0.5772$. The extreme-value regression model with dispersion covariates is defined by (1) and by two systematic components given by

$$g(\mu_t) = \eta_t = \eta(x_t, \beta)$$

and

$$h(\phi_t) = \delta_t = \delta(z_t, \gamma),$$

where $\beta = (\beta_1, \ldots, \beta_k)^\top$ and $\gamma = (\gamma_1, \ldots, \gamma_m)^\top$ are vectors of unknown regression parameters ($\beta \in \mathbb{R}^k$ and $\gamma \in \mathbb{R}^m$) and $x_t$ and $z_t$ are observations on $k$ and $m$ covariates ($k < n$ and $m < n$). Here, $\eta(\cdot, \cdot)$ and $\delta(\cdot, \cdot)$ are continuously twice differentiable (possibly nonlinear) functions in the second argument. Finally, $g(\cdot)$ and $h(\cdot)$ are known strictly monotonic and twice differentiable link functions that maps $\mathbb{R}$ and $\mathbb{R}^+$, respectively. Let $X$ be the derivative matrix of $\eta_t = (\eta_1_t, \ldots, \eta_n_t)^\top$ with respect to $\beta^\top$. Analogously, let $Z$ be the derivative matrix of $\delta = (\delta_1, \ldots, \delta_n)^\top$ with respect to $\gamma^\top$. It is assumed that $\text{rank}(X) = k$ and $\text{rank}(Z) = m$ for all $\beta$ and $\gamma$.

Let $\ell(\theta)$ be the log-likelihood function of the model defined by (1)-(3) given the vector of observations $y = (y_1, \ldots, y_n)$. We have

$$\ell(\theta) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi_t),$$

where

$$\ell_t(\mu_t, \phi_t) = -\log(\phi_t) - \frac{y_t - \mu_t}{\phi_t} - \exp \left( - \frac{y_t - \mu_t}{\phi_t} \right),$$

with $\mu_t$ and $\phi_t$ defined so that (2) and (3) hold. The score function, obtained by differentiating the log-likelihood function with respect to the unknown parameters, is denoted by $U = (U_\beta(\beta, \gamma)^\top, U_\gamma(\beta, \gamma)^\top)^\top$ and $I$ and $J$ denote the expected and the observed information matrices.

### 3 Main result and discussion

Let $\theta = (\beta^\top, \gamma^\top)^\top$ be the unknown parameter vector that indexes the extreme-value regression model (1)-(3). In what follows, $\nu = (\nu_1, \ldots, \nu_r)^\top$ represents the parameter of interest and $\psi = (\psi_1, \ldots, \psi_s)^\top$ is the nuisance parameter; notice that $r + s = k + m$. We consider likelihood-based tests of the null hypothesis $\mathcal{H}_0 : \nu = \nu_0$, where $\nu_0$ is a fixed $r$-vector. Clearly, such tests may be inverted to give confidence sets for $\nu$. Also, let $J_{\psi\psi}$ denote the $s \times s$ observed information matrix corresponding to $\psi$. Similarly, $A_{\psi\psi}$ denotes a matrix formed from the $(r + s) \times (r + s)$ matrix $A$ by dropping the rows and columns that correspond to the interest parameter. Additionally, hat and tilde indicate evaluation at the unrestricted ($\hat{\theta}$) and at the restricted ($\tilde{\theta}$) maximum likelihood estimator of $\theta$ under $\mathcal{H}_0$, respectively. For instance, $\hat{I} = I(\hat{\theta}), \tilde{I} = I(\tilde{\theta})$ and $\hat{J} = J(\hat{\theta})$. 

by inserting \( \hat{\theta} \) for \( \theta_1 \) and \( \tilde{\theta} \) for \( \theta \) after the expected values are computed. Notice that \( \overline{q} \) is a \((r+s)\)-vector and \( \Upsilon \) is a \((r+s) \times (r+s)\) matrix. Under \( \mathcal{H}_0 \), \( w \) is distributed as \( \chi^2_r \) with error of order \( n^{-1} \) while \( w^* \) follows this distribution with high degree of accuracy (Skovgaard, 2001, p. 7). Simulation results in Ferrari and Cysneiros (2008) and Ferrari and Pinheiro (2010) suggest that tests that use \( w^* \) are much less size distorted than those that are based on \( w \).

In order to obtain the adjusted likelihood ratio statistic (4) in the extreme-value regression model (1)-(3), one needs to obtain score vector, the observed and expected information matrices, \( J \) and \( I \) respectively, the vector \( q \) and the matrix \( \Upsilon \). We obtained

\[
U_{\beta}(\beta, \gamma) = X^\top \Phi^{-1}T(1 - z^\top)
\]

\[
U_{\gamma}(\beta, \gamma) = Z^\top \Phi^{-1}H(z - ZZ^\top - 1),
\]

\[
J = \begin{bmatrix} J_{\beta\beta} & J_{\beta\gamma} \\ J_{\gamma\beta} & J_{\gamma\gamma} \end{bmatrix}
\]

with

\[
J_{\beta\beta} = X^\top \Phi^{-1}T \left( Z^\top \Phi^{-1} + (I - Z^\top)ST \right) TX - \left[ t^\top (I - Z^\top) T \Phi^{-1} \right] \hat{X},
\]

\[
J_{\beta\gamma} = J_{\gamma\beta}^\top = X^\top \Phi^{-1}T(I - Z^\top + ZZ^\top)H \Phi^{-1}Z
\]

\[
J_{\gamma\gamma} = Z^\top \Phi^{-1}H \left( (-I + 2Z - 2ZZ^\top + ZZ^\top) \Phi^{-1} + (-I + Z^\top - ZZ^\top) QH \right) HZ + \left[ t^\top (I - Z + ZZ^\top) H \Phi^{-1} \right] \hat{Z},
\]

\[
I = \begin{bmatrix} I_{\beta\beta} & I_{\beta\gamma} \\ I_{\gamma\beta} & I_{\gamma\gamma} \end{bmatrix}
\]

with

\[
I_{\beta\beta} = X^\top \Phi^{-2}T^2X,
\]

\[
I_{\beta\gamma} = I_{\gamma\beta}^\top = (E - 1)X^\top \Phi^{-1}TH \Phi^{-1}Z
\]

\[
I_{\gamma\gamma} = (1 + \Gamma(2))Z^\top \Phi^{-1}H^2 \Phi^{-1}Z,
\]
\[
\begin{array}{c}
\bar{q} = \begin{bmatrix}
\hat{X}^\top \hat{\Phi}^{-1} \hat{T} (I - MD) \hat{t} \\
\hat{Z}^\top \hat{\Phi}^{-1} \hat{H} (C(EI + ND) - I) \hat{t}
\end{bmatrix}
\end{array}
\]

and
\[
\bar{T} = \begin{bmatrix}
\hat{X}^\top \hat{\Phi}^{-1} \hat{T} CMD^{-\frac{1}{2}} \hat{\Phi}^{-\frac{1}{2}} \hat{X} \\
\hat{Z}^\top \hat{\Phi}^{-1} \hat{H} CND^{-\frac{1}{2}} \hat{\Phi}^{-\frac{1}{2}} \hat{X} \\
\end{bmatrix}
\]

\( t \) is the \( n \)-dimensional column vector of ones, \( z = (z_1, \ldots, z_n)^\top \) with \( z_t = (y_t - \mu_t)/\phi_t \), \( z^* = (\exp(-z_1), \ldots, \exp(-z_n))^\top \), \( Z = \text{diag}(z_1, \ldots, z_n) \), \( Z^\dagger = \text{diag}(\exp(-z_1), \ldots, \exp(-z_n)) \), \( \Phi = \text{diag}(\phi_1, \ldots, \phi_n) \), \( T = \text{diag}(1/g'(\mu_1), \ldots, 1/g'(\mu_n)) \), \( H = \text{diag}(1/h'(\phi_1), \ldots, 1/h'(\phi_n)) \), \( S = \text{diag}(g''(\mu_1), \ldots, g''(\mu_n)) \), \( Q = \text{diag}(h''(\phi_1), \ldots, h''(\phi_n)) \), \( \hat{X} = \partial^2 \eta / \partial \beta \partial \eta \) and \( \hat{Z} = \partial^2 \delta / \partial \eta \partial \gamma \), \( n \times k \times k \) and \( n \times m \times m \) arrays, respectively, \( C = \text{diag}(\phi_{11}/\phi_1, \ldots, \phi_{1n}/\phi_n) \), \( D = \text{diag}(\mu_{11} - \mu_1/\phi_1, \ldots, \mu_{1n} - \mu_n/\phi_n) \), \( D^\dagger = \text{diag}(\exp(-\mu_{11} - \mu_1/\phi_1), \ldots, \exp(-\mu_{1n} - \mu_n/\phi_n)) \), \( M = \text{diag}(\Gamma(1 + \phi_{11}/\phi_1), \ldots, \Gamma(1 + \phi_{1n}/\phi_n)) \), \( N = \text{diag}(\Gamma(1) (1 + \phi_{11}/\phi_1), \ldots, \Gamma(1) (1 + \phi_{1n}/\phi_n)) \), \( P = \text{diag}(\Gamma(2) (1 + \phi_{11}/\phi_1), \ldots, \Gamma(2) (1 + \phi_{1n}/\phi_n)) \), \( \Gamma(\cdot) \) denotes the gamma function and \( \Gamma(1)(\cdot) \) and \( \Gamma(2)(\cdot) \) are its first and second derivatives, respectively.

We present Monte Carlo simulation results on the small sample behaviour of the likelihood ratio test \( (w) \), the Wald test \( (W) \), the score test \( (S_R) \), the gradient test \( (S_T) \), and the adjusted likelihood ratio test \( (w^*) \). The Wald, score and gradient statistics are given by \( W = (\bar{\nu} - \nu_0)^\top (\hat{P}^\nu)^{-1} (\bar{\nu} - \nu_0) \), \( S_R = \bar{U}_y^\top \bar{P}^\nu \bar{U}_y \) and \( S_T = \bar{U}_y^\top (\bar{\nu} - \nu_0) \). Notice that the gradient statistic is very simple to compute since it does not involve the information matrix, neither observed nor expected.

We now consider model (1) with constant dispersion and systematic component for the location parameter given by
\[
\mu_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2},
\]

We consider the null hypothesis, namely \( H_0 : \beta_2 = 0 \) (\( r = 1 \)) to be tested against two-sided alternatives. We set \( \beta_0 = 1 \), \( \beta_1 = 1 \) and \( \beta_2 = 0 \). The covariate values are chosen as random draws from the \( \mathcal{U}(0, 1) \) distribution, \( \phi \) is set at 1.105171, and the samples sizes are 15, 20, 30 and 40. The number of Monte Carlo replications is 10,000 and the nominal levels of the tests are \( \alpha = 10\% \), 5\% and 1\%. All simulations were carried out using the Ox matrix programming language (Doornik, 2009). Table 1 contains the null rejection rates of the tests.

| \( n \) | \( \alpha = 10\% \) & | \( \alpha = 5\% \) & | \( \alpha = 1\% \) & |
|-------|------|------|------|------|
|       |     |      |      |      |
| 15    | 16.6 | 22.2 | 10.2 | 13.4 | 10.0 |
| 20    | 14.1 | 19.2 | 9.5  | 11.9 | 9.9  |
| 30    | 12.5 | 16.3 | 9.3  | 11.1 | 9.7  |
| 40    | 12.2 | 15.1 | 9.8  | 11.2 | 10.3 |

Table 1: Null rejection rates (%)
4 Conclusion

Our simulation results suggest that the likelihood ratio and the Wald tests can be markedly oversized in small and moderate-sized samples. The gradient test can be oversized, but much less than the two aforementioned tests. The score test is even less size distorted and can be conservative in some cases. The adjusted likelihood ratio test obtained in this paper performs better than all the others. Although it requires more computational effort, it is the least size distorted in most cases and it is, therefore, recommended for practical applications. We shall emphasize that our simulations were carried out in extreme-value regression models with linear and non-linear predictors for both location and dispersion parameters. All simulation results exhibited reasonably similar behavior.

References


