Comparison of exact and approximate simultaneous confidence regions in nonlinear regression models

Confronto tra la regione di confidenza esatta ed approssimata nei modelli di regressione nonlineari

Claudia Furlan and Cinzia Mortarino

Abstract Accuracy measures for parameter estimates represent a tricky issue in nonlinear models. Practitioners often use the separate marginal confidence intervals for each parameter. However, these can be extremely misleading due to the curvature of the parameter space of the nonlinear model. For low parameter dimensions, routines for evaluating approximate simultaneous confidence regions are available in the most common software programs, but the degree of accuracy also depends on the intrinsic nonlinearity of the model. In this paper, the accuracy of the marginal confidence intervals, Hartley's exact simultaneous confidence region (sCR), and the most widespread approximate sCR are compared via both real data and simulations, for discrete time diffusion models in the class of nonlinear regression models.

Abstract *Nei modelli non lineari non è scontato riuscire ad ottenere misure di accuratezza delle stime. Nella pratica spesso si usano gli intervalli di confidenza marginali per ogni parametro, ma questa procedura può portare a risultati inaffidabili a causa della curvatura dello spazio parametrico tipico dei modelli non lineari. Nei più comuni software si può trovare implementato il calcolo della regione di confidenza simultanea approssimata per un numero ridotto di parametri, ma il livello di copertura esatto dipende dal grado di non linearità intrinseca del modello. In questo lavoro, nell'ambito dei modelli di regressione non lineari e in particolare per i modelli di diffusione a tempo discreto, si confrontano fra loro i livelli di copertura degli intervalli di confidenza marginali, della regione di confidenza simultanea (sCR) esatta di Hartley e della sCR approssimata più utilizzata.*

Key words: nonlinear models, simultaneous confidence region, Hartley's simultaneous confidence region, Bass model

1 Introduction

Nonlinear models are the natural modelling framework for many real-world phenomena. Unlike linear models, accuracy measures for parameter estimates, such as confidence intervals or confidence regions, may represent a difficult task due to the intrinsic curvature of the parameter space. A common mistake is relying on

Claudia Furlan e-mail: furlan@stat.unipd.it · Cinzia Mortarino e-mail: mortarino@stat.unipd.it
Department of Statistical Sciences, University of Padova, Italy.

marginal confidence intervals, whose use can be misleading. Simultaneous confidence regions are usually available in the most commonly used software programs, only in the approximate form at least for low parameter dimensions.

The problem of constructing exact confidence regions for the parameters of nonlinear models has received little attention in the past (Lee et al, 2002), since this is computationally intensive. Given the complexity of obtaining an exact simultaneous confidence region (sCR), a few approximations have been proposed (Seber and Wild, 1989) under the normality assumption of homoscedastic errors. For instance, the so-called 'approximate' sCR is derived by approximating the nonlinear model via a linear Taylor expansion, thereby taking advantage of the asymptotic normality of the estimator. Thus, the approximate confidence levels of the 'approximate' sCR are valid asymptotically. This approximation is computationally more attractive, since it corresponds to hyperellipsoids.

Under the normality assumption of homoscedastic errors, Hartley (1964) proposed an exact sCR based on inverting an exact test. However, the power of the exact test, and thus the coverage probability of the corresponding sCR, depends on the choice of the idempotent projection matrix. More recently, Demidenko (2017) studied the exact statistical properties in small samples.

Among the nonlinear regression models, in this paper, we focus on two of the most widespread discrete time diffusion models of products and technologies; these are the Bass model (BM) and the Generalized Bass model (GBM) which have three and six parameters, respectively. We analyze two case studies based on real data, namely Algerian natural gas production and Austrian solar thermal capacity. In this paper we derive and compare Hartley (1964)'s exact sCR with Guseo (1983)'s projection matrix with the 'approximate' sCR, in terms of accuracy, via simulation studies. The simulation studies are performed to explore the effect of increasing the parameter dimension with different model structures. Specifically, the BM is considered in a constrained version and in its full form (two and three parameters respectively), while the GBM is considered with two different intervention functions (six parameters in both versions). The simulation studies are performed for lifecycles with the same diffusion characteristics as those of the two case studies.

2 Diffusion models

Let us denote with n the number of observations used to fit the model and with y the n-dimensional vector obtained by stacking the observed y_i values, $i = 1,2,\ldots,n$. Similarly, $f(\vartheta;t)$ will denote the vector $f(\vartheta;t) = \{f(\vartheta;t_1), f(\vartheta;t_2), \ldots, f(\vartheta;t_n)\}'$.

For a general nonlinear regression model,

$$y_i = f(\vartheta;t_i) + \varepsilon_i, \quad \vartheta \in \mathbb{R}^k, \quad i = 1,2,\ldots,n, \tag{1}$$

where $\mathfrak{C}(\Theta) \subset \mathbb{R}^k$ is the exact sCR with confidence level $1-\alpha$.

In the application of diffusion models, the starting point is given by an observed time series reporting sales data or consumption/production of a technological innovation. In this work, we consider the BM and GBM (Bass et al, 1994). Let $z(t)$ be the cumulative data, at time t, and $w(t)$ an intervention function. The GBM is:

$$z(t) = m\frac{1-e^{-(p+q)\int_0^t w(\tau)d\tau}}{1+\frac{q}{p}e^{-(p+q)\int_0^t w(\tau)d\tau}}, \tag{2}$$

where m is the market potential, p is the innovation coefficient, q is the imitation coefficient, and $w(t)$ can be any integrable function. Below, we examine the model arising when $w(t)$ is specified by the so-called *exponential* shock (Guseo and Dalla Valle, 2005),

$$w(t) = 1 + c_1 e^{b_1(t-a_1)} I_{t\geq a_1}, \tag{3}$$

which allows us to describe the diffusion of a product for which, at time a_1, we observe a rapid and reversible ($b_1 < 0$) shock with intensity c_1; and when $w(t)$ is specified by the so-called *rectangular* shock,

$$w(t) = 1 + c_1 I_{a_1\leq t\leq b_1}, \tag{4}$$

which allows us to describe the diffusion of a product for which we observe a constant shock with intensity c_1, in the time interval $[a_1, b_1]$. We will denote the GBM of Eq. (2) with $w(t)$ as in (3) by GBM$_{exp}$, and with $w(t)$ as in (4) by GBM$_{rect}$. These structures are a special case of model (1), where $f(\vartheta;t)$ is represented by $z(t)$ in Eq. (2), with $w(t)$ specified as in (3) or (4). The cumulative time series data (y) can easily be used to jointly estimate all the parameters (m, p, q, a_1, b_1, c_1) of the model using nonlinear least squares. Finally, the BM is the special case of the GBM, when $w(t) = 1$, $t \in \mathbb{R}^+$.

3 Exact and approximate inference

Hartley (1964) proposed a method for evaluating an exact sCR. This method gives the possibility of verifying whether any point in the parameter space Θ belongs to the exact $(1-\alpha)$ level sCR. Hartley (1964)'s exact sCR is

$$\mathfrak{C}(\Theta) = \left\{ \vartheta : \frac{[y-f(\vartheta;t)]'P[y-f(\vartheta;t)]}{[y-f(\vartheta;t)]'[I_n-P][y-f(\vartheta;t)]} \leq \frac{k}{n-k}F_{1-\alpha}(k,n-k) \right\}, \tag{5}$$

where $F_{1-\alpha}(k,n-k)$ is the $1-\alpha$ percentile of a Snedecor's F distribution with k and $n-k$ degrees of freedom. The confidence level is exact if ε's components can be assumed to be independent and distributed according to a Gaussian distribution. In this work, we assume error homoscedasticity with variance σ^2, and we use the projection matrix proposed by Guseo (1983), $P = F(F'F)^{-1}F'$, where F denotes the $n \times k$ matrix obtained by deriving the vector $f(\vartheta;t)$ with respect to the k-dimensional vector ϑ, $F = \frac{\partial f(\vartheta;t)}{\partial \vartheta}$. This assumption about P leads to the exact sCR that is used in this work:

$$\mathfrak{C}(\Theta) = \left\{ \vartheta : \frac{[y-f(\vartheta;t)]'F(F'F)^{-1}F'[y-f(\vartheta;t)]}{[y-f(\vartheta;t)]'[I_n-F(F'F)^{-1}F'][y-f(\vartheta;t)]} \leq \frac{k}{n-k}F_{1-\alpha}(k,n-k) \right\}. \tag{6}$$

The so-called 'approximate' sCR, denoted here by $\mathfrak{I}(\Theta)$, is derived by approximating the nonlinear model by a linear Taylor expansion, taking advantage of the asymptotic normality of the estimator. The 'approximate' sCR, $\mathfrak{I}(\Theta) \subset \mathbb{R}^k$, is

$$\mathfrak{I}(\Theta) = \left\{ \vartheta : (\vartheta - \hat{\vartheta})' \hat{F}' \hat{F} (\vartheta - \hat{\vartheta}) \leq k s^2 F_{1-\alpha}(k, n-k) \right\}, \tag{7}$$

where $\hat{F} = F(\hat{\vartheta})$, and s^2 is the sample variance. As the linear approximation is valid asymptotically, $\mathfrak{I}(\Theta)$ will have the correct confidence level of $1-\alpha$, asymptotically.

To evaluate $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$, we derived the expression of the components of F for the GBM$_{exp}$, GBM$_{rect}$, and BM, but we omit them here for brevity.

4 Real data analysis

One field that is currently under the public eye is the diffusion of renewable and nonrenewable energy systems. One energy system of each type was chosen in this study, namely Algerian natural gas production, in billion cubic metres (BCM), with annual data from 1970 to 2004 ($n = 35$, source: www.bp.com), and Austrian thermal solar capacity, in MW$_{th}$, with annual data from 1982 to 2008 ($n = 27$, source: www.estif.org). The data are shown in Figure 1.

We compare the $\mathfrak{C}(\Theta)$ of Eq. (6) and $\mathfrak{I}(\Theta)$ of Eq. (7) in both time series, for increasing parameter dimensions. To accomplish this, we have selected three nested models: the constrained BM (with m fixed, thus $k = 2$), the BM ($k = 3$), and GBM ($k = 6$). In this paper, we decide to show $\mathfrak{I}(\Theta)$ and $\mathfrak{C}(\Theta)$ only for the BM $k = 3$. To do that, we used a grid of 12,190,801 points. Each point in the grid has been subsequently tested to assess inclusion in $\mathfrak{C}(\Theta)$ or $\mathfrak{I}(\Theta)$, via conditions (6) and (7), respectively, with $1-\alpha = 0.95$. For the natural gas production, the proportion of common points with respect to $\mathfrak{C}(\Theta)$ is 0.706, while it is 0.7 for $\mathfrak{I}(\Theta)$. For the solar thermal capacity, the proportion of common points with respect to $\mathfrak{C}(\Theta)$ is 0.442, while it is 0.512 for $\mathfrak{I}(\Theta)$. The degree of overlap is smaller in both energy systems compared with what we found with $k = 2$. The representation of these points, for both time series, is shown in Figure 2, together with the representation of a grid covering the parallelepiped generated by combining the marginal confidence intervals of level 0.95, evaluated separately with the Bonferroni method for the three parameters. For both time series, the difference between $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$ is larger than that observed for the case with $k = 2$. In particular, for the solar thermal capacity, the discrepancy between $\mathfrak{I}(\Theta)$ and $\mathfrak{C}(\Theta)$ is much bigger, and the shape of $\mathfrak{C}(\Theta)$ is far from being ellipsoidal.

Moving to the case with $k = 6$, we fitted the GBM$_{exp}$ to the natural gas production and the GBM$_{rect}$ to the solar thermal capacity. Only the fitted values when $k = 6$ are plotted in Figure 1, since the GBMs were found to be the best models, according to the F-test for nested models. For the natural gas production, the proportion of common points with respect to $\mathfrak{C}(\Theta)$ is 0.060, while it is 0.103 for $\mathfrak{I}(\Theta)$. For the solar thermal capacity, the proportion of common points with respect to $\mathfrak{C}(\Theta)$ is 0.333, while it is 0.984 for $\mathfrak{I}(\Theta)$. The degree of overlap is drastically low, denoting a high curvature of the space $f(\vartheta; t)$. Especially in this final case, $\mathfrak{I}(\Theta)$ appears to

be extremely small with respect to $\mathfrak{C}(\Theta)$, thereby excluding many values that belong to $\mathfrak{C}(\Theta)$.

5 Simulation study

In this section, $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$ are compared in terms of coverage probability, for the models used in Section 4: the constrained BM, BM, and GBM. For each model, we generated N=1,000 simulated time series, using estimates as true values. Moreover, the length of simulated time series corresponds to the number of real data ($n = 35$ for the natural gas production and $n = 27$ for the solar thermal capacity). In this way, the simulation study investigates the coverage probability of $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$, with data with diffusion characteristics, intervention functions, and stage of the lifecycle corresponding to those of the time series considered in Section 4.

Given each simulated time series, $j = 1,\ldots,N$, we evaluated parameter estimates and both sCRs, $\mathfrak{C}(\Theta)_j$ and $\mathfrak{I}(\Theta)_j$. We then tested whether the true values of the parameters used to generate the N time series were included in $\mathfrak{C}(\Theta)_j$ or $\mathfrak{I}(\Theta)_j$. The proportion of $\mathfrak{C}(\Theta)_j$ and $\mathfrak{I}(\Theta)_j$ containing the true values represents the coverage probability. The coverage probabilities of $\mathfrak{I}(\Theta)$ and $\mathfrak{C}(\Theta)$ for the constrained BM ($k = 2$), BM ($k = 3$), and GBM ($k = 6$) are plotted in Figure 3: the difference between $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$ increases with the model complexity, and it is negligible for $k = 2$ for both energy systems. It emerges that the coverage probability of $\mathfrak{C}(\Theta)$ also decreases as the variability $(\sigma/m)^2$ increases, but its decay is less severe than what happens with $\mathfrak{I}(\Theta)$. This is especially true for $k = 6$. In summary, for both the case studies, the degree of overlap of $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$ decreased as k increased, denoting that the parameter space curvature increased with k, as well as that the shape of $\mathfrak{C}(\Theta)$ was progressively farther from being ellipsoidal. We could conclude that $\mathfrak{I}(\Theta)$ can be satisfactorily used with a low parameter dimension, or with a moderate parameter dimension only if the variability is limited. The validity of this result is limited to data with diffusion characteristics similar to those used in this paper. Further research is required to generalize the effect of the lifecycle stage on the coverage probability.

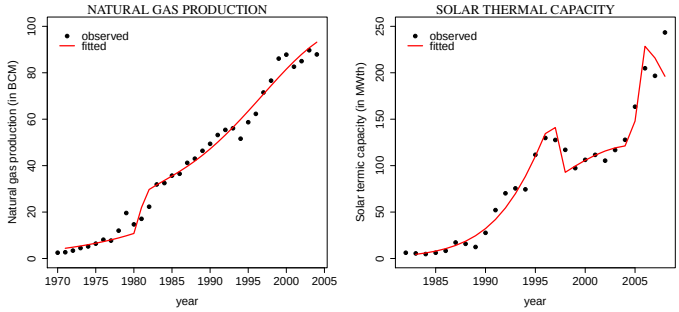

Fig. 1 The lines denote the fitted values (GBM_{exp} and GBM_{rect}, respectively).

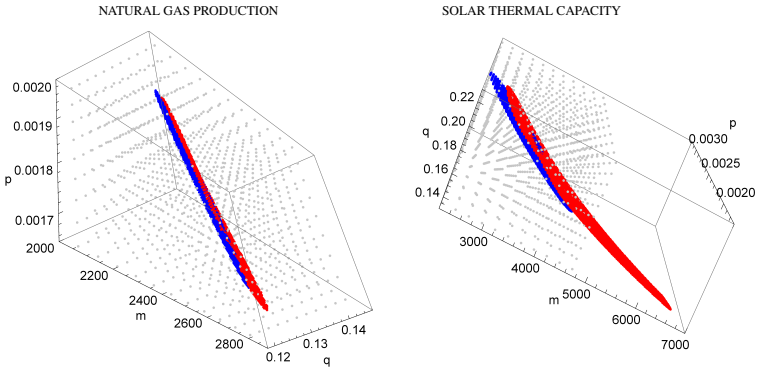

Fig. 2 BM ($k = 3$, $1 - \alpha = 0.95$). $\mathfrak{C}(\Theta)$ is in red and $\mathfrak{I}(\Theta)$ in blue. Grey points represent sCIs.

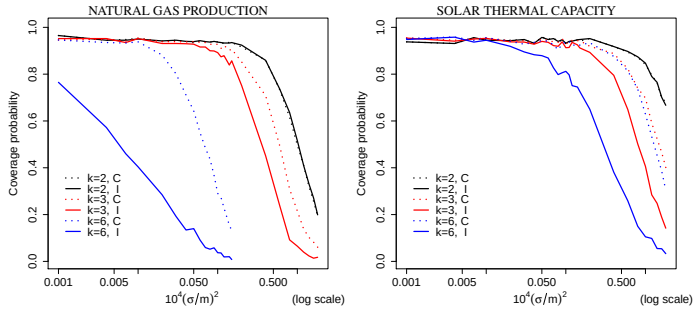

Fig. 3 Coverage probability of $\mathfrak{C}(\Theta)$ and $\mathfrak{I}(\Theta)$. Values of $(\sigma/m)^2$ are in the log scale.

References

Bass FM, Krishnan TV, Jain DC (1994) Why the Bass model fits without decision variables. Marketing Science 13(3):203–223

Demidenko E (2017) Exact and approximate statistical inference for nonlinear regression and the estimating equation approach. Scandinavian Journal of Statistics 44(3):636–665

Guseo R (1983) Confidence regions in non linear regression. Proceedings of the 44th Session of International Statistical Institute, Madrid, 12-22/9/83 pp 333–336

Guseo R, Dalla Valle A (2005) Oil and gas depletion: diffusion models and forecasting under strategic intervention. Statistical Methods and Applications 14(3):375–387

Hartley HO (1964) Exact confidence regions for the parameters in non-linear regression laws. Biometrika 51(3/4):347–353

Lee A, Nyangoma S, Seber G (2002) Confidence regions for multinomial parameters. Computational Statistics & Data Analysis 39(3):329–342

Seber G, Wild C (1989) Nonlinear Regression. Wiley: New York