

Robust inference in presence of unsuspected correlation

Inferenza robusta rispetto a non specificate forme di dipendenza

Luca Greco¹ and Laura Ventura²

¹ Dipartimento di Statistica, Università Ca' Foscari di Venezia (***)
e-mail: greco@dst.unive.it

² Dipartimento di Scienze Statistiche, Università di Padova (†)
e-mail: ventura@stat.unipd.it

Riassunto: Nell'ambito dei modelli lineari, la non corretta specificazione della struttura di dipendenza degli errori ha effetti negativi sull'inferenza basata sugli M-stimatori. In particolare, nell'ambito dei modelli lineari, gli intervalli di confidenza per un coefficiente scalare di regressione, basati sulla funzione di quasi-verosimiglianza, possono risultare meno ampi di quanto dovrebbero. Obiettivo di questo lavoro è proporre delle procedure che permettano la costruzione di funzioni di quasi-verosimiglianza che tengano conto dell'eventuale presenza di correlazione nei dati.

Keywords: AR, Bootstrap, Correlated errors, M-estimators, Quasi-profile likelihood.

1. Introduction

Unsuspected correlation may arise when observations are collected in sequence or on a spatial domain. For instance, this is the case of environmental, geophysical, meteorological, agricultural, physical, economic data, as it has been acknowledged by many scientists in different research areas. The problem is also outlined in Hampel *et al.* (1986, section 8.1).

The presence of unsuspected dependence structures among data may have non negligible effects on M-estimation procedures in linear regression models, affecting Wald, score and quasi-likelihood test statistics. Even when data are weakly correlated, the coverage of confidence intervals obtained under the assumption of independence may be quite different from the nominal level, as pointed out by Field and Wiens (1994) and Field and Zhou (2003).

The aim of this paper is to obtain robust confidence intervals for regression coefficients based on pseudo-likelihood functions allowing for the possibility of dependence among data. These pseudo-likelihood functions should lead to reliable inference for both uncorrelated and weakly correlated data, in order to reveal useful when one is unwilling or unable to justify either the assumption of independence or any particular model of dependence.

The inferential procedures we are interested in making robust against small departures from the assumption of uncorrelated data are those based on the quasi-likelihood function (Adimari and Ventura, 2002). Investigation is restricted to the use of Huber's Proposal II for regression models but results are quite general. In Section 2 robust-against-

(***) Corresponding author

(†) Partially supported by a grant from MIUR, Italy.

dependence techniques are described and illustrative examples are provided in Section 3. Concluding remarks follow in Section 4.

2. Robust-against-dependence quasi-profile likelihood

Consider a linear regression model of the form $y = \beta^T \varepsilon$, where β is a p -dimensional vector of coefficients and ε 's are normal distributed with mean 0, variance σ^2 and may be correlated. In general, an M-estimate for $\theta = \beta/\sigma$ is defined as the root of an unbiased estimating equation $\psi(\theta; y) = 0$. In this context, we focus on Huber estimation of the regression coefficients, with scale parameter estimated by Huber's Proposal II.

Without loss of generality, suppose we are interested in drawing inference about a scalar regression coefficient $\tau = \beta_1 \leq \beta_2 \leq \dots \leq \beta_p$. The vector parameter θ can be thought to be partitioned as $\theta = (\tau, \lambda)^T$, where λ is treated as a $(p-1)$ -dimensional nuisance parameter. The estimating equation $\psi(\theta; y)$ is similarly partitioned as $\psi(\theta; y) = (\psi_\tau(\tau, \lambda; y), \psi_\lambda(\tau, \lambda; y))^T$, where $\psi_\tau(\tau, \lambda; y)$ and $\psi_\lambda(\tau, \lambda; y)$ are the estimating functions corresponding to τ and λ , respectively. Let λ_τ denote the estimate derived from ψ_λ when τ is considered as known, i.e. the solution of $\psi_\lambda(\tau, \lambda; y) = 0$. When we desire inference about τ to be based on a pseudo-likelihood function, instead on classical Wald-type statistics, a quasi-profile loglikelihood for τ can be considered (Adimari and Ventura, 2002), that is

$$\ell_{QP}(\tau) = \sum_{t=1}^n w_t \psi_\tau(\tau, \lambda_t; y_t) \quad (1)$$

where τ_0 is an arbitrary constant and

$$\omega(\tau, \lambda_\tau) = \frac{\frac{\partial}{\partial \tau} \psi_\tau(\tau, \lambda_\tau) - \psi_\tau(\tau, \lambda_\tau) \frac{\partial}{\partial \tau} \lambda_\tau}{\psi_\tau(\tau, \lambda_\tau)} \quad (2)$$

The scale adjustment (2) is obtained so that the rescaled profile estimating function $\omega(\tau, \lambda_\tau) \psi_\tau(\tau, \lambda_\tau)$ has bias and information bias of order $O(n^{-1})$, as for the ordinary profile score function. In practice, (2) is necessary to obtain quasi-profile loglikelihood-type tests based on (1) with the standard χ^2_1 asymptotic distribution.

In simple models, such as linear models, (2) can be computed by first-order analytical approximations to the moments of the derivatives of λ_τ . In general, it is preferable to resort to a simulation process in which the needed quantities are estimated by parametric bootstrap sampling from the normal model (Bellio *et al.*, 2005).

When data are correlated, the estimating function ψ_τ is still unbiased with respect to any symmetric model. The misspecification of the errors' dependence structure can, however, lead to wrong inferences. If we ignore the correlations between the data, standard errors result to be underestimated, significance test will lead to statistics that are too large and p -values that are too small, leading to rejecting more often than we should. For this reason, to take into account unsuspected correlation, quasi-likelihood based inferential procedures may be obtained from the same estimating function but involving some dependence structures in the bootstrap scheme when evaluating (2).

To this end, we consider two methods to overcome this hindrance. In the first method, the algorithm named **BP** described in Field and Zhou (2003) is implemented in the

bootstrap evaluation of the scale adjustment $\omega_\tau = \lambda_\tau$. The error process is approximated by an AR model. For each τ on a fixed grid, we compute λ_τ , we estimate the order and the parameters ϕ_s of the AR model from the centered constrained residuals r_s . Then, new residuals from the AR fit are obtained as $u = \phi_1 r_{-1} - \phi_p r_{-p}$ and, finally, after centering the u 's, bootstrap resampling is applied to them. These bootstrap samples are, then, used to evaluate (2).

The second method, **TAR**, is based on a testing procedure and on an AR approximation of the error process. This choice is supported by several numerical studies (Gallant, 1987) and appears to be quite robust against dependence process misspecification. A linear model assuming uncorrelated errors is fitted and its residuals are tested for autocorrelation by the Durbin-Watson statistic. An estimate of the AR parameter is $\phi = \frac{\sum_{r=2}^n r r_{-1}}{\sum_{r=1}^n r^2}$. If the test is significant at a level chosen to be α , then, for each fixed τ , a parametric bootstrap is performed by simulating from a linear model with errors following an AR process with autoregressive parameter set equal to ϕ .

3. Examples

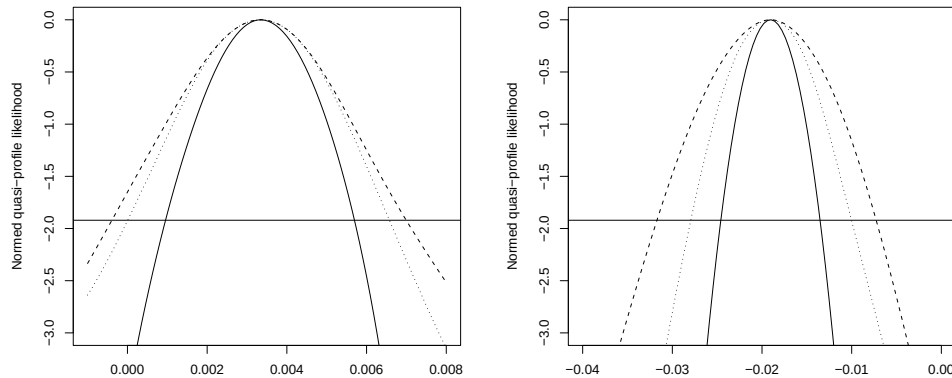
The performance of the proposed method is verified by two applications to real data. The considered data sets refer to situations in which it is known that dependence processes among the errors are operating: in time, in the first example, and in the spatial setting, for what concern the second example. The goal is to prove that our methods, effectively, correct the inferential conclusions based on the quasi-profile likelihood, developed under the assumption of no correlation, from an unbiased estimating equation defining an M-estimator.

Icecream consumption data. This data set is taken from Hand *et al.* (1982). The icecream consumption, weekly household income, price and average temperature have been measured in 30 consecutive four-weeks periods from the 18th of March 1951 to the 1st of July 1953. Icecream consumption is taken as the response variable. Assume we desire inference on the coefficient of the covariate weekly household income. Inspection of the ordinary least squares residuals shows positive correlation but we ignore it and do not make any change in the specified linear model.

Wheat yield data. We consider the famous wheat-yield data of Mercer and Hall (1911), used by several authors to show and fit spatial dependence. The wheat yields are from $10 \text{ feet} \times 10 \text{ feet}$ plots. The number of plots is 250 and they are arranged in a lattice with 20 rows in the East-West direction and 25 columns in the North-south direction. We consider a standard linear model with covariates corresponding to row and column coordinates. It is well known that the residuals from this fit show significant spatial dependence. Anyway, suppose we are interested in drawing inference on the column coordinate effect without specifying any spatial structure on the lattice.

In these two examples both the **BP** and the **TAR** quasi-likelihood function allow to quantify the consequences of the misspecification of the dependence structure. Actually, they provide a useful display of the effect of accounting for correlation, resulting in larger confidence intervals, as plotted in Figure 1. Bootstrap computations were applied with 500 repetitions.

Figure 1: Icecream data (left) and Wheat yield data (right): Quasi-profile (solid), BP (dashed), TAR (dotted). The horizontal line gives the .95 confidence interval.



4. Concluding remarks

In this paper we present two methods to deal with the problem of unsuspected correlation in supposedly independent and identically distributed data. The procedures are applied to the construction of a robust-against-dependence quasi-profile likelihood, for inference on a scalar regression parameter based on M-estimators. The proposed methods are shown to take into account the effect of this kind of misspecification. However, the properties of these techniques need more investigation with regard to both their finite sample and asymptotic behavior.

References

- Adimari, G., Ventura, L. (2002) Quasi-profile loglikelihood for unbiased estimating functions, *Annals of the Institute of Statistical Mathematics*, 54, 235–244.
- Bellio, R., Greco, L., Ventura, L. (2005) Adjusted quasi-profile likelihood from estimating functions, *Working Paper Series*, 10, Department of Statistics, University of Padua. Submitted.
- Field C., Wiens D. (1994) One step M-estimators in the linear model, with dependent errors, *The Canadian Journal of Statistics*, 22, 219–231.
- Field C., Zhou J. (2003) Confidence intervals based on robust regression, *Journal of Statistical Planning and Inference*, 115, 425–439.
- Gallant, A.R. (1987) *Nonlinear Statistical Model*, Wiley, New York.
- Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. (1986) *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.
- Hand D.J., Daly F., Lunn A.D., McConway K.J., Ostrowski E. (1994), *Small Data Sets*, Chapman Hall.
- Mercer W.B. and Hall A.D. (1911) The experimental error of field trials, *Journal of Agricultural Science*, 4, 107–132.