Available online at www.sciencedirect.com
ScienceDirect
Automatica 44 (2008) 407–417
automatica
www.elsevier.com/locate/automatica

Robust estimation in multiple linear regression model with non-Gaussian noise☆

Ayşen D. Akkaya*, Moti L. Tiku[1]

Department of Statistics, Middle East Technical University, 06531 Ankara, Turkey

Received 3 April 2006; received in revised form 15 May 2007; accepted 13 June 2007
Available online 11 December 2007

Abstract

The traditional least squares estimators used in multiple linear regression model are very sensitive to design anomalies. To rectify the situation we propose a reparametrization of the model. We derive modified maximum likelihood estimators and show that they are robust and considerably more efficient than the least squares estimators besides being insensitive to moderate design anomalies.

Keywords: Linear regression; Robustness; Data anomaly; Modified maximum likelihood; Outliers

1. Introduction

The motivation for this paper comes from Puthenpura and Sinha (1986) who give examples in the context of off-line identification. They show that the commonly used least square estimators (LSE) of the parameters in a multiple linear regression model are not efficient if the data is very noisy. They also show that the maximum likelihood equations are very sensitive to gross errors (outliers) and have convergence problems. As pointed out by Puthenpura and Sinha (1986, p. 231), outliers in data can occur due to large disturbances, data transmission errors, failures in transducers and A/D converters, etc. They work out modified maximum likelihood estimators (MMLE) from data replicated at each design point and show that the estimators are robust to outliers, a very desirable property. See also Aström (1980), Sayed, Nascimento, and Capparrone (2002) and Subramanian and Sayed (2004) who reflect on the importance of robustness. In working out their estimators, Puthenpura and Sinha censor a certain proportion of extreme observations as in Tiku (1978). However, replications at each design point are often not available. Also, the estimators of σ based on censored samples can have substantial downward bias (Tiku, 1980). We work out MMLE from complete samples when the noise is non-Gaussian and only one observation is available at each design point as in most situations. We show that our estimators are efficient and robust to outliers and other data anomalies.

The methodology of modified maximum likelihood estimation originated with Tiku (1967, 1989) and Tiku and Suresh (1992) and has been used extensively (Puthenpura & Sinha, 1986; Schneider, 1986; Tiku & Akkaya, 2004; Tan & Tabatabai, 1988; Tiku, Tan, & Balakrishnan, 1986; Vaughan, 2002). Another difficulty with the LSE is that their variances are profoundly influenced by the design values x_{ij} $(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant q)$ and the fact is that the design values are not always pre-determined. If there are outliers (Tiku, 1977) in the design values, the diagonal elements in $(\mathbf{X}'\mathbf{X})^{-1}$ will be small and the LSE of the regression coefficients θ_j $(1 \leqslant j \leqslant q)$ will appear to be efficient. If there are inliers (Akkaya & Tiku, 2005) in the design values, the diagonal elements in $(\mathbf{X}'\mathbf{X})^{-1}$ will be large and the LSE will appear to be inefficient. This is easier to see when $q = 1$. To rectify the situation, we suggest a reparametrization of the model. We show that the LSE for this model are insensitive to design anomalies. We derive the MMLE and show that they are robust (in particular to outliers) and considerably more efficient than the LSE besides being insensitive to design anomalies.

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor Torsten Söderström.
* Corresponding author.
E-mail address: akkay@metu.edu.tr (A.D. Akkaya).
[1] Present address. McMaster University, Ont., Canada.

0005-1098/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2007.06.029

2. Reparametrized model and estimation

The multiple linear regression model we propose is

$$y_i = \theta_0 + \sum_{j=1}^{q} \theta_j u_{ij} + e_i \quad (1 \leqslant i \leqslant n, \;\; 1 \leqslant j \leqslant q), \tag{1}$$

where

$$u_{ij} = (x_{ij} - \bar{x}_j)/s_j, \quad \bar{x}_j = (1/n)\sum_{i=1}^{n} x_{ij}$$

and

$$s_j^2 = (1/n)\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2.$$

We assume, of course, that no two x_j are constant multiples of one another. The model is a natural analogue of the conditional expectation

$$E(Y|X_1{=}x_1, \ldots, X_q{=}x_q){=}\, \mu_y + \sum_{j=1}^{q} \delta_j \sigma_y (x_j - \mu_j)\,/\sigma_j \tag{2}$$

in numerous elliptically contoured multivariate distributions, e.g., multivariate normal. Since in (1), x_{ij} $(1 \leqslant j \leqslant q)$ are all nonstochastic, μ_j and σ_j in (2) are replaced by $\bar{x}_j$ and s_j, respectively. That gives the model (1).

Least squares estimators: Assuming that e_i are iid with mean zero and variance σ^2, the LSE are

$$\tilde{\theta}_0 = \bar{y}, \quad \tilde{\boldsymbol{\theta}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{Y} \quad \text{and} \quad \tilde{\sigma} = s_e, \tag{3}$$

$$s_e^2 = \sum_{i=1}^{n}\left(y_i - \bar{y} - \sum_{j=1}^{q}\tilde{\theta}_j u_{ij}\right)^2 \Big/ (n - q - 1),$$

$$\bar{y} = (1/n)\sum_{i=1}^{n} y_i.$$

Like $\tilde{\theta}_0$ and $\tilde{\sigma}$, $\tilde{\theta}_j$ $(1 \leqslant j \leqslant q)$ are invariant to location and scale of x_{ij} $(1 \leqslant i \leqslant n)$, i.e., if some or all x_{ij} are replaced by $a_j + b_j x_{ij}$ $(1 \leqslant i \leqslant n)$, the values of $\tilde{\theta}_j$ do not change. The variance–covariance matrix of $(\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_q)$ is

$$Cov(\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_q) = (\mathbf{U}'\mathbf{U})^{-1}\sigma^2,$$
$$\sum_{i=1}^{n} u_{ij}^2 = n \quad (1 \leqslant j \leqslant q). \tag{4}$$

The LSE $\tilde{\theta}_0$ and $\tilde{\sigma}$ are uncorrelated with $\tilde{\theta}_j$ $(1 \leqslant j \leqslant q)$, and

$$V(\tilde{\theta}_0) = \frac{\sigma^2}{n} \quad \text{and} \quad V(\tilde{\sigma}) \cong \frac{\sigma^2}{2n}\left(1 + \frac{1}{2}\lambda_4\right),$$

$\lambda_4{=}(\mu_4/\mu_2^2){-}3$ (Roy & Tiku, 1962). For Gaussian noise, $\lambda_4{=}0$. If the distribution of e is symmetric, $\tilde{\theta}_0$ is uncorrelated with $\tilde{\sigma}$.

Numerous engineering applications of least squares estimation and variants of it are given in Sinha and Kuszta (1983) and Kailath, Sayed, and Hassibi (2000).

3. Symmetric distributions

Suppose that the noise has one of the distributions in the family

$$f(e) = \frac{\Gamma(p/2)}{\sigma\sqrt{k}\,\Gamma(1/2)\Gamma(p - 1/2)}\left\{1 + \frac{e^2}{k\sigma^2}\right\}^{-p}, \quad -\infty < e < \infty, \tag{5}$$

$k = 2p - 3$ and $p \geqslant 2$. Note that $E(e) = 0$ and $V(e) = \sigma^2$. The distribution of $t = \sqrt{(\upsilon/k)}(e/\sigma)$ is Student's t with $\upsilon = 2p - 1$ degrees of freedom.

The likelihood function is

$$L \propto \left(\frac{1}{\sigma}\right)^n \prod_{i=1}^{n}\left[1 + \frac{(\mathbf{Y} - \mathbf{1}\theta_0 - \mathbf{U}\theta)'(\mathbf{Y} - \mathbf{1}\theta_0 - \mathbf{U}\theta)}{k\sigma^2}\right]^{-p}.$$

The MLE are solutions of the maximum likelihood equations

$$\partial \ln L/\partial\theta_0 = 0,$$
$$\partial \ln L/\partial\theta_j = 0 \quad (1 \leqslant j \leqslant q)$$

and

$$\partial \ln L/\partial\sigma = 0. \tag{6}$$

The equations are expressions in terms of the intractable functions

$$g(z_i) = z_i/\{1 + (1/k)z_i^2\}, \quad z_i = e_i/\sigma \;\; (1 \leqslant i \leqslant n)$$

and have no explicit solutions. Solving the $q + 2$ nonlinear equations (6) by iteration is a very difficult task and there are convergence problems as said earlier. Therefore, we work out MMLE that are known to be asymptotically equivalent to MLE (Bhattacharyya, 1985; Vaughan & Tiku, 2000, Appendix A). For small n, they are known to be essentially as efficient as MLE and the two are numerically very close to one another (Schneider, 1986, p. 104; Tiku & Vaughan, 1997, pp. 890–892; Tiku et al., 1986, pp. 101, 106–107; Vaughan, 2002, p. 228).

The MMLE are obtained in three steps: (i) the equations in (6) are expressed in terms of the ordered variates $z_{(i)} = e_{(i)}/\sigma$ (accomplished simply by replacing z_i by $z_{(i)}$), (ii) the functions $g(z_{(i)})$ are replaced by linear approximations such that the differences between the two converge to zero as n tends to infinity and (iii) the resulting equations (called modified maximum likelihood equations) are solved. The solutions (called MMLE) are explicit functions of the concomitant observations

$$(y_{[i]}, u_{[i]1}, \ldots, u_{[i]q}) \quad (1 \leqslant i \leqslant n),$$

i.e., the vector of observations $(y_i, u_{i1}, \ldots, u_{iq})$ associated with the ith ordered (in order of increasing magnitude) residual $e_{(i)} = \sigma z_{(i)}$:

$$e_{(i)} = y_{[i]} - \theta_0 - \sum_{j=1}^{q} \theta_j u_{[i]j} \quad (1 \leqslant i \leqslant n). \tag{7}$$

Now, consider the linear approximations

$$g(z_{(i)}) \cong \alpha_i + \beta_i z_{(i)}, \quad 1 \leqslant i \leqslant n. \tag{8}$$