SENSITIVITY ANALYSIS OF M-ESTIMATES OF NONLINEAR REGRESSION MODEL: INFLUENCE OF DATA SUBSETS*

JAN ÁMOS VIŠEK

Department of Stochastic Informatics, Institute of Information Theory and Automation,
Academy of Sciences, the Czech Republic

Department of Macroeconomics, Institute of Economic Studies, Charles University,
Karlova 5, CZ-110 00 Prague 1, the Czech Republic, e-mail: visek@mbox.fsv.cuni.cz

(Received November 15, 1998; revised September 18, 2000)

Abstract. Asymptotic representations of the difference of $M$-estimators of parameters of nonlinear regression model for the full data and for the subsample of data are given for the following three situations: i) fix number of points excluded from data, ii) increasing number, however asymptotically negligible part of data excluded, and finally iii) asymptotically fix portion of data excluded. Asymptotic normality of the difference of estimators (for the two former cases) is proved.

Key words and phrases: Sensitivity analysis, asymptotic representation, asymptotic distribution of the difference of $M$-estimators, testing subsample stability of estimates, scale invariance, diagnostics of subsets of influential points, diversity of estimates.

1. Introduction

An analysis of the influence of individual datum or of data subsets on the results of any data-processing procedure has an eminent importance for the applications. Therefore in any theory oriented on data-processing, a part of it has been always devoted to this topic. In regression analysis one may find an amount of references to the papers treating this problem for instance in the monographs about the sensitivity analysis of estimation by Atkinson (1985), Belsley et al. (1980), Chatterjee and Hadi (1988) or Rousseeuw and Leroy (1987), to give at least some of them.

In the linear regression there is a well-known formula for the difference of the least squares estimators for the full data set and for a subset of data containing $n-1$ observations, namely

$$\hat{\beta}_{LS}^{(n-1,\ell)} - \hat{\beta}_{LS}^{(n)} = -\left( [X^{(n-1,\ell)}]^T X^{(n-1,\ell)} \right)^{-1} X_{\ell} Y_{\ell} - X_{\ell}^T \hat{\beta}_{LS}^{(n)}$$

where notation is nearly selfexplaining, nevertheless, $X^{(n-1,\ell)}$ is the design matrix after deletion of the $\ell$-th row from the full design matrix $X$ and $X_{\ell}$ is the $\ell$-th row (considered as a column vector) of the design matrix for the full data (see e.g. Chatterjee and Hadi (1988), Višek (1992a) or Žvára (1989)).

The present paper derives, in a form of asymptotic representations of Bahadur type, analogical formulas for the $M$-estimators of nonlinear regression model for the three situations, namely when a fix number of observations is excluded from the data set, when an

*Research was supported by grant of GA UK number 255/2000/A EK /FSV.
increasing but asymptotically negligible number of observations is excluded and finally, when an asymptotically fix portion of observations is excluded. The representations for the two latter cases hint the asymptotic normality of the respective difference, and so they allow to establish a test of subsample stability of estimates in the sense of Víšek (1992a). We shall see later that it may help to select the most adequate model for given data in the case when various methods give considerably different estimates.

Let us start with notations.

Notations. Let $N$ denote the set of all positive integers, $R$ the real line, $R^+$ its positive part and $(\Omega, B, P)$ a probability space. We shall consider for all $n \in N$ the nonlinear regression model

$$Y_i = g(X_i, \beta^0) + e_i, \quad i = 1, 2, \ldots, n$$

where $\{X_n\}_{n=1}^\infty$ is a fix sequence of vectors from $R^q$, $\beta^0 \in R^p$ and $\{e_i\}_{i=1}^\infty$, $e_n : \Omega \to R$ is a sequence of independent and identically distributed random variables (i.i.d.r.v.) with $Ee_1 = 0$ and $Ee_i^2 = \sigma^2 \in (0, \infty)$. $F(z)$ will denote the distribution function (d.f.) of $e_i$, $\sigma^{-1}$, respectively. Finally, having denoted $I_k = \{i_1, i_2, \ldots, i_k; 1 \leq i_1 < i_2 < \ldots < i_k\} \subset N$, let us define

$$\hat{\beta}(n) = \arg \min_{\beta \in R^p} \sum_{i=1}^n \rho([Y_i - g(X_i, \beta)]\hat{\sigma}^{-1})$$

and

$$\hat{\beta}(n, I_k) = \arg \min_{\beta \in R^p} \sum_{i \in \{1, 2, \ldots, n\} \setminus I_k} \rho([Y_i - g(X_i, \beta)]\hat{\sigma}^{-1})$$

where $\rho$ is an absolutely continuous function (with a derivative $\psi$) and $\hat{\sigma}$ is a preliminary estimator of the scale of residuals. $\hat{\sigma}$ is assumed to be regression-invariant and scale-equivariant in order to achieve regression- and scale-equivariance of $\hat{\beta}(n)$ (see in Bickel (1975) or Jurečková and Sen (1993) and condition C.iii below). To simplify all considerations we will assume that the same estimate of the scale will be used for the full data set and for the “reduced” data. This assumption does not represent substantial restriction of generality either from the theoretical point of view or for the applications. In former case it only burden the notations by some additional items and prolongs the proofs of theorems. In the latter case it asks for the employment of robust estimators of scale and for a check whether after deletion the change of scale estimate is not dramatic. If however instability of the scale estimate occurs, we may expect the same instability of the estimate of regression model. In such a case we shall probably prefer another estimator of regression model anyway, see discussion below. Let us give now conditions under which we will derive the results.

2. Conditions

**CONDITIONS A.** (i) There is a positive $\delta_0$ such that for any $\beta \in R^p$, $\|\beta - \beta^0\| < \delta_0$

$$\frac{\partial}{\partial \beta_j} g(x, \beta) \quad (j = 1, 2, \ldots, p) \quad \text{and} \quad \frac{\partial^2}{\partial \beta_j \partial \beta_k} g(x, \beta) \quad (j, k = 1, 2, \ldots, p)$$

exist for any $x \in \{X_n\}_{n=1}^\infty$. Let us denote the vector of the first partial derivative and the matrix of the second derivatives simply by $g'(x, \beta)$ and $g''(x, \beta)$, respectively, and their coordinates and elements by $g'_j(x, \beta)$ and $g''_{jk}(x, \beta)$. 