Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 779–802
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Sensitivity analysis of constrained linear L_1 regression: perturbations to response and predictor variables

Mingren Shi[a], Mark A. Lukas[b,*]

[a] *Mathematics and Computing, University of Southern Queensland, Toowoomba, Qld 4350, Australia*

[b] *Mathematics and Statistics, Murdoch University, South Street, Murdoch, WA 6150, Australia*

Received 20 March 2003; received in revised form 26 March 2004; accepted 27 March 2004

Abstract

The active set framework of the reduced gradient algorithm is used to develop a direct sensitivity analysis of linear L_1 (least absolute deviations) regression with linear equality and inequality constraints on the parameters. We investigate the effect on the L_1 regression estimate of a perturbation to the values of the response or predictor variables. For observations with nonzero residuals, we find intervals for the values of the variables for which the estimate is unchanged. For observations with zero residuals, we find the change in the estimate due to a small perturbation to the variable value. The results provide practical diagnostic formulae. They quantify some robustness properties of constrained L_1 regression and show that it is stable, but not uniformly stable. The level of sensitivity to perturbations depends on the degree of collinearity in the model and, for predictor variables, also on how close the estimate is to being nonunique. The results are illustrated with numerical simulations on examples including curve fitting and derivative estimation using trigonometric series.

MSC: 62J05; 90C31; 65K05

Keywords: Sensitivity analysis; Stability; L_1 regression; Least absolute deviations; Robustness; Diagnostics; Active set; Reduced gradient algorithm

* Corresponding author. Tel.: 61893602423; fax: 61893606332.

E-mail addresses: shi@usq.edu.au (M. Shi), M.Lukas@murdoch.edu.au (M.A. Lukas).

doi:10.1016/j.csda.2004.03.015

1. Introduction

Consider a linear model $\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where $\mathbf{y}$ is an $n\times 1$ response vector corresponding to the $n\times p$ design matrix X of predictor variable values, $\boldsymbol{\beta}$ is an unknown $p\times 1$ vector of parameters and $\boldsymbol{\varepsilon}$ is an $n\times 1$ vector of random errors. For our purposes it will be convenient to write the model as a system of linear equations $y_i=\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}+\varepsilon_i$, $i=1,\dots,n$, where $\mathbf{x}_i^{\mathrm{T}}$ is the ith row of X.

In many applications there are additional linear constraints that must be satisfied by some or all of the parameters, for example, positivity. In particular, biometric and econometric models of the form $\bar{y}_i=\bar{\beta}_1\bar{x}_{i2}^{\beta_2}\bar{x}_{i3}^{\beta_3}e^{\varepsilon_i}$, with positive $\bar{\beta}_1$, β_2 and β_3, are of this type after a logarithmic transformation (see p. 444 in Judge et al., 1985). Constrained regression problems also arise naturally in the important areas of parametric (and nonparametric) curve and surface fitting, and in the estimation of solutions of ill-posed and inverse problems from noisy data (see Wahba, 1990). Here, extra information such as the value of the solution at some point leads to a linear equality constraint on the parameters. Extra information such as positivity, monotonicity, concavity or convexity of the solution leads to a set of linear inequality constraints on the parameters (see Wahba, 1982 and O'Leary and Rust, 1986). For notational simplicity we will write the linear equality constraints as $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i=0$, $i\in\mathscr{E}=\{n+1,\dots,n+n_{\mathscr{E}}\}$, and the inequality constraints as $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i\leqslant 0$, $i\in\mathscr{I}=\{n+n_{\mathscr{E}}+1,\dots,n+n_{\mathscr{E}}+n_{\mathscr{I}}\}$.

In the unconstrained case, it is usual to estimate $\boldsymbol{\beta}$ using least squares (L_2) regression. For the constrained problem, restricted or constrained least squares regression (as well as other approaches) have been used (see Knautz, 2000). However, as is well known, the least squares method is not robust; it is not optimal for error distributions with long tails and the estimates are overly sensitive to outliers.

Over the past 25 years there has been growing interest in the method of least absolute deviations or L_1 regression as an alternative to least squares regression. For the linear model with linear constraints above, the L_1 regression estimate of $\boldsymbol{\beta}$ is the solution to the problem (denoted LL_1)

$$\text{minimize}\quad S(\boldsymbol{\beta})=\sum_{i=1}^{n}|\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i|,\qquad \boldsymbol{\beta}\in\Re^p, \tag{1.1A}$$

$$\text{subject to}\quad \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i=0,\qquad i\in\mathscr{E}=\{n+1,\dots,n+n_{\mathscr{E}}\}, \tag{1.1B}$$

$$\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}-y_i\leqslant 0,\qquad i\in\mathscr{I}=\{n+n_{\mathscr{E}}+1,\dots,n+n_{\mathscr{E}}+n_{\mathscr{I}}\}, \tag{1.1C}$$

where we assume that $n_{\mathscr{E}}<p<n+n_{\mathscr{E}}+n_{\mathscr{I}}$.

An important advantage of L_1 regression over L_2 regression is its robustness. For the unconstrained problem (denoted UL_1), it is well known that the L_1 regression estimator can resist a few large errors in the data $\mathbf{y}$. In fact (see Lemma 3.1 or Bloomfield and Steiger, 1983), the optimal solution (regression estimate) to UL_1 is completely unaffected by a perturbation of $\mathbf{y}$ that maintains the same signs of the residuals. Bloomfield and Steiger (1983, Section 2.3) also derived the generalized influence function for L_1 regression, which shows its robustness with respect to y_i,