Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 1213 – 1231
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Sensitivity analysis of constrained linear L_1 regression: Perturbations to constraints, addition and deletion of observations

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

[a]*Mathematics and Statistics, Murdoch University, Murdoch, WA 6150, Australia*
[b]*Mathematics and Computing, University of Southern Queensland, Toowoomba, Qld 4350, Australia*

Received 20 May 2003; received in revised form 26 March 2004; accepted 6 April 2004
Available online 20 July 2006

Abstract

This paper extends the direct sensitivity analysis of Shi and Lukas [2005, Sensitivity analysis of constrained linear L_1 regression: perturbations to response and predictor variables. Comput. Statist. Data Anal. 48, 779–802] of linear L_1 (least absolute deviations) regression with linear equality and inequality constraints on the parameters. Using the same active set framework of the reduced gradient algorithm (RGA), we investigate the effect on the L_1 regression estimate of small perturbations to the constraints (constants and coefficients). It is shown that the constrained estimate is stable, but not uniformly stable, and in certain cases it is unchanged. We also consider the effect of addition and deletion of observations and determine conditions under which the estimate is unchanged. The results demonstrate the robustness of L_1 regression and provide useful diagnostic information about the influence of observations. Results characterizing the (possibly non-unique) solution set are also given. The sensitivity results are illustrated with numerical simulations on the problem of derivative estimation under a concavity constraint.

MSC: 62J05; 90C31; 65K05

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

1. Introduction

Consider a linear model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i$, $i = 1, \ldots, n$, where the $p \times 1$ parameter vector $\boldsymbol{\beta}$ is known to satisfy additional linear equality constraints $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i = 0$ and/or inequality constraints $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i \leqslant 0$. Important applications occur in parametric (and non-parametric) curve and surface fitting, and in the estimation of solutions of ill-posed and inverse problems from noisy data (see Wahba, 1990). In many such problems there is some extra information about the unknown curve or solution that can be used to constrain the parameters. In particular, if the solution (or some linear functional, e.g., an integral) is known to have a certain fixed value at some point, we obtain an equality constraint on the parameters. If the solution is known to be positive, monotone, concave or convex, this leads to certain linear inequality constraints on the parameters (see Wahba, 1982; and O'Leary and Rust, 1986). Constrained regression problems also arise in certain biometric and econometric models (see Judge et al., 1985).

* Corresponding author. Tel.: +61 8 93602423; fax: +61 8 93606332.
E-mail addresses: M.Lukas@murdoch.edu.au (M.A. Lukas), shi@usq.edu.au (M. Shi).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2004.04.014

Using the method of least absolute deviations, we define the constrained L_1 regression estimate of $\boldsymbol{\beta}$ to be the solution of the problem (denoted LL_1):

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

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

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

where $\mathscr{E}$ and $\mathscr{I}$ refer to equalities and inequalities, respectively, and we assume that $n_{\mathscr{E}} < p < n + n_{\mathscr{E}} + n_{\mathscr{I}}$. Shi and Lukas (2005) investigated the sensitivity of the solution to (1.1) with respect to perturbations in the responses y_i and row vectors $\mathbf{x}_i^{\mathrm{T}}$, $1 \leqslant i \leqslant n$, of the design matrix X. The analysis was done using the active set framework of the reduced gradient algorithm (RGA) developed in Shi and Lukas (2002). In this paper we use the same framework and extend the sensitivity analysis to cover perturbations to the constraints (both the constants and coefficients) and the addition and deletion of observations.

It is well known that unconstrained L_1 regression is robust; the L_1 estimate is resistant to outliers in the response $\mathbf{y}$ and may not be changed at all by some large perturbations in $\mathbf{y}$. The same is true for constrained L_1 regression, and the estimate is stable, but not uniformly stable, with respect to small perturbations in $\mathbf{y}$ and $\mathbf{x}_i$, $i = 1, \ldots, n$ (Shi and Lukas, 2005; Ellis, 1998). Here we show that similar results apply for perturbations to the constraints in (1.1); the constrained estimate is stable for small perturbations and it may not even change with some large perturbations in the constraints. The stability is not uniform but depends on the degree of ill-conditioning and, for the constraint coefficient vectors, also on how close the estimate is to being non-unique.

For unconstrained L_1 regression, the effect of deletion of observations has been investigated by Narula and Wellington (1985), Dupačová (1992) and Castillo et al. (2001). These works are based on the LP formulation of the L_1 regression problem. Here we use the direct active set approach on the general LL_1 problem (1.1) to find conditions under which an existing solution remains optimal after the deletion of an observation. We also show how the RGA can be used to find a new solution to the problem with a deleted observation, starting from the original solution. The L_1 version of the Cook distance can then be computed efficiently to determine the influence of each observation on the L_1 regression estimate. This provides important diagnostic information about the model. A result about the addition of an observation shows that constrained L_1 regression is robust with respect to the new response.

A brief description of the RGA framework and optimality conditions for (1.1) is given in Section 2. It is known that the solution of an L_1 regression problem can be non-unique (Bloomfield and Steiger, 1983). In Section 3 we derive results characterizing the solution set of (1.1) and show that it can be computed efficiently using the RGA. In Sections 4 and 5, we investigate the effect on the solution to (1.1) of a perturbation to the constants and coefficients of the constraints, respectively. Results relating to the deletion and addition of an observation are derived in Sections 6 and 7, respectively.

In Section 8 we consider the problem of estimating the derivative $f(t) = g'(t)$ from data $y_i = g(t_i) + \varepsilon_i$, $i = 1, \ldots, n$, under the concavity constraint $f''(t) \leqslant 0$. The estimation of derivatives arises in many applications, in particular in the analysis of growth curves (Gasser et al., 1984; Eubank, 1988) and pharmacokinetic response curves (Song et al., 1995). Numerical differentiation is an ill-posed problem, meaning that the solution is sensitive to errors in the data (see Anderssen and Bloomfield, 1974), and such problems were a major motivation for this work. We use a truncated trigonometric series for the estimate $f_p(t)$ of the derivative and find the estimate by solving a constrained linear L_1 regression problem. Results of numerical simulations illustrate the sensitivity results from Sections 4–6, and also a result from Shi and Lukas (2005) on perturbations to the responses. We also consider the effect of increasing the number of points at which the constraint $f_p''(t) \leqslant 0$ is evaluated.

2. Active set framework

We use the same notation, definitions and optimality results as in Shi and Lukas (2005), but we repeat them here for completeness (see also Bloomfield and Steiger, 1983; Osborne, 1985). An inequality constraint $\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i \leqslant 0$ is said to be active at $\boldsymbol{\beta}$ if $\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - y_i = 0$ (equality constraints are automatically active). Similarly we say that a model