Statistics and Probability Letters 83 (2013) 829–835

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Does adding data always improve linear regression estimates?

A.V. den Boer*

Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 5 September 2012
Received in revised form 3 December 2012
Accepted 3 December 2012
Available online 10 December 2012

Keywords:
Least-squares linear regression
Strong consistency
Inconsistency

ABSTRACT

Intuitively one might expect that the quality of statistical estimates cannot worsen if they are based on more data. We show in a least-squares linear regression setting that this intuition is wrong. Adding data may worsen the quality of parameter estimates, and in fact may even cause a design sequence to lose strong consistency.

1. Introduction and motivation

Consider a linear model of the form

$$y_t = x_t^T \theta^{(0)} + \epsilon_t, \quad (t \in \mathbb{N}), \tag{1}$$

where $y_t \in \mathbb{R}$ is a response variable (sometimes called observation, output, or measurement), $x_t \in \mathbb{R}^d$ an explanatory variable (sometimes called input or design variable), ϵ_t a zero-mean random variable denoting a disturbance term or measurement error, and $\theta^{(0)} \in \mathbb{R}^d$ an unknown parameter; with v^T we denote the transpose of a vector v. It is common practice to estimate the value of the unknown parameter $\theta^{(0)}$ by least-squares linear regression: given input variables $x_1, \ldots, x_t$ and observations $y_1, \ldots, y_t$, this estimate is defined as $\hat{\theta}_t = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^t (y_i - x_i^T \theta)^2$.

In many applications, notably control problems, the design variables x_t are not given a priori, but have to be determined by a decision maker. An important question is how $(x_t)_{t \in \mathbb{N}}$ should be chosen such that $\hat{\theta}_t$ is strongly consistent, meaning that $\hat{\theta}_t$ converges a.s. to $\theta^{(0)}$ as t grows large.

Lai et al. (1979) establish strong consistency of $\hat{\theta}_t$ under the condition

$$\lim_{t \to \infty} \left(\sum_{i=1}^t x_i x_i^T \right)^{-1} = 0, \tag{2}$$

when the design sequence $(x_t)_{t \in \mathbb{N}}$ is deterministic, the disturbance terms $(\epsilon_t)_{t \in \mathbb{N}}$ form a martingale difference sequence with respect to a filtration $\{\mathcal{F}_t\}_{t \in \mathbb{N}}$, and $\sup_{t \in \mathbb{N}} E[\epsilon_t^2 \mid \mathcal{F}_{t-1}] < \infty$. In addition, they show that (2) is necessary for strong consistency if the error terms $(\epsilon_t)_{t \in \mathbb{N}}$ are i.i.d. and $\sum_{i=1}^t x_i x_i^T$ is invertible for some $t \in \mathbb{N}$.

* Tel.: +31 0 40 247 5647.
E-mail address: a.v.d.boer@tue.nl.

0167-7152/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2012.12.001

If the design is not deterministic, (2) is not strong enough to guarantee strong consistency. Lai and Wei (1982) show that if x_t is $\mathcal{F}_{t-1}$-measurable for all $t \in \mathbb{N}$, and $\sup_{t\in\mathbb{N}} E[|\epsilon_t|^{2+\delta} \mid \mathcal{F}_{t-1}] < \infty$ a.s. for some $\delta > 0$, then $\hat{\theta}_t$ converges a.s. to $\theta^{(0)}$ if (2) holds a.s. and

$$\lim_{t\to\infty} \frac{\lambda_{\min}\left(\sum_{i=1}^{t} x_i x_i^T\right)}{\log\left(\lambda_{\max}\left(\sum_{i=1}^{t} x_i x_i^T\right)\right)} = \infty \quad \text{a.s.}, \tag{3}$$

where $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ denote the smallest and the largest eigenvalue of a symmetric matrix A. They also give an example where strong consistency does not hold, and where the left-hand side of (3) converges a.s. to a random variable. This implies that for designs $(x_t)_{t\in\mathbb{N}}$ that satisfy (2) but violate (3) strong consistency can in general only be concluded if the design is deterministic.

Now, suppose a design $(x_t)_{t\in\mathbb{N}}$ is predictable w.r.t. $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ (i.e. x_t is $\mathcal{F}_{t-1}$-measurable for all $t \in \mathbb{N}$), and $(x_t)_{t\in\mathbb{N}}$ contains a *deterministic* subsequence $(x_{t_j})_{j\in\mathbb{N}}$ such that

$$\lim_{j\to\infty} \left(\sum_{i=1}^{j} x_{t_i} x_{t_i}^T\right)^{-1} = 0. \tag{4}$$

Intuitively one might expect that adding data can never worsen the quality of a least-squares estimate. Since the least-squares estimate that only uses data $(x_{t_j}, y_{t_j})_{j\in\mathbb{N}}$ is strongly consistent, this would imply that the estimate $\hat{\theta}_t$ based on *all* data $(x_t, y_t)_{t\in\mathbb{N}}$ is strongly consistent as well. In general, this would imply that strong consistency follows if (2) is satisfied on a deterministic subsequence of the design sequence. These considerations motivate us to study the question whether indeed adding data can never worsen the quality of $\hat{\theta}_t$.

A second motivation comes from sequential decision problems under uncertainty, where a decision maker simultaneously has to estimate unknown parameters and maximize a reward function that depends on these parameters. Broder and Rusmevichientong (2012) and den Boer and Zwart (submitted for publication) are examples from the dynamic pricing literature that study such problems: the decisions $x_t = (1, p_t)^T \in \mathbb{R}^2$ correspond to selling prices p_t that have to be determined by a firm, the output y_t corresponds to observed demand in time period $t \in \mathbb{N}$, and the objective of the firm is to maximize the cumulative expected revenue $\sum_{t=1}^{T} E[y_t p_t]$ in T time periods. The challenge in these problems is that each selling price p_t does not only influence the immediately earned revenue, but also the quality of future parameter estimates which influence the revenues earned in the future. Broder and Rusmevichientong (2012) analyze a pricing policy (called "MLE-CYCLE") in which estimates are formed based on a deterministic design sequence. They show numerically that a similar pricing policy (called "MLE-CYCLE-S") that uses *all* available data to form estimates has a better performance, but they do not provide a mathematical justification. Their numerical results nevertheless seem to confirm the intuition that adding data can only be beneficial.

In this brief paper we show that this intuition is not true. We consider the linear model (1), and measure the quality of the least-squares estimate $\hat{\theta}_t$ by $E[\|\hat{\theta}_t - \theta^{(0)}\|^2]$. In Proposition 1, Section 2, we provide a sufficient condition on x_{t+1} such that

$$E[\|\hat{\theta}_{t+1} - \theta^{(0)}\|^2] > E[\|\hat{\theta}_t - \theta^{(0)}\|^2].$$

We illustrate the condition for the simple, widely used linear model $y_t = \theta_0^{(0)} + \theta_1^{(0)} z_t + \epsilon_t$.

Proposition 2, Section 3, shows that the deterioration of parameter estimates caused by adding data can, in a qualitative sense, be quite bad: we provide an example of a design sequence for which the corresponding sequence of least-squares estimators is strongly consistent, and show that this design sequence can be augmented by extra data points such that the resulting sequence of least-squares estimators is *not* strongly consistent anymore.

2. Increasing expected estimation error by adding data

Consider the regression model (1). Assume that the error terms $(\epsilon_t)_{t\in\mathbb{N}}$ form a martingale difference sequence w.r.t. a filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$,

$$\inf_{t\in\mathbb{N}} E\left[\epsilon_{t+1}^2 \mid \mathcal{F}_t\right] > 0 \quad \text{a.s.},$$

and suppose x_t is $\mathcal{F}_{t-1}$-measurable, for all $t \in \mathbb{N}$.

The least-squares linear regression estimate $\hat{\theta}_t$ of $\theta^{(0)}$, based on $(x_i, y_i)_{1\le i\le t}$, is equal to

$$\hat{\theta}_t = \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{t} (y_i - x_i^T\theta)^2.$$