Computational Statistics and Data Analysis 54 (2010) 2899–2913

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Globally robust confidence intervals for simple linear regression

Jorge Adrover[a], Matias Salibian-Barrera[b,*]

[a] Universidad Nacional de Córdoba, CONICET and CIEM, Argentina
[b] University of British Columbia, Canada

ARTICLE INFO

Article history:
Received 13 August 2008
Received in revised form 1 May 2009
Accepted 7 May 2009
Available online 18 May 2009

ABSTRACT

It is well known that when the data may contain outliers or other departures from the assumed model, classical inference methods can be seriously affected and yield confidence levels much lower than the nominal ones. This paper proposes robust confidence intervals and tests for the parameters of the simple linear regression model that maintain their coverage and significance level, respectively, over whole contamination neighbourhoods. This approach can be used with any consistent regression estimator for which maximum bias curves are tabulated, and thus it is more widely applicable than previous proposals in the literature. Although the results regarding the coverage level of these confidence intervals are asymptotic in nature, simulation studies suggest that these robust inference procedures work well for small samples, and compare very favourably with earlier proposals in the literature.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Consider the simple linear regression model where we observe a bivariate random sample $(Y_1, X_1), \ldots, (Y_n, X_n)$ satisfying

$$Y_i = \beta_0 + \beta_1 (X_i - \mu_X) + \sigma_0 \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where X_i are univariate explanatory variables, $\mu_X = \text{med}(X_i)$, the errors ϵ_i follow a known distribution F_0 and satisfy $\text{med}(\epsilon_i|X_i) = 0$, $i = 1, \ldots, n$. In general, one assumes that the data are generated by a distribution $H_{\boldsymbol{\theta}}$ belonging to a parametric family of distributions $\{H_{\boldsymbol{\theta}}\}$, with $\boldsymbol{\theta} \in \mathbb{R}^2$. To allow for outliers and other departures from the model, we will assume that the data follow a distribution H in an ϵ-contamination neighbourhood $\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})$ of the true underlying parametric model. More specifically,

$$\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}}) = \left\{H = (1-\varepsilon)\, H_{\boldsymbol{\theta}} + \varepsilon H^*,\, H^* \text{ an arbitrary distribution on } \mathbb{R}^2\right\}, \tag{2}$$

where $0 < \varepsilon < 0.5$.

Confidence intervals based on maximum likelihood estimators may be seriously affected by a small proportion of atypical observations (see, e.g. Tukey and McLaughlin (1963), Dixon and Tukey (1968), Huber (1968, 1970), Barnett and Lewis (1994), Fraiman et al. (2001) and Adrover et al. (2004)). We will say that a confidence interval is robust if it is able to maintain a high coverage level and a reasonable length when the data comes from any distribution in the contamination neighbourhood (2). Formally, we have the following:

Definition 1. A confidence interval (L_n, U_n) for $\theta \in \mathbb{R}$ is called *globally robust of level* $(1-\alpha)$ if it satisfies the following conditions:

[*] Corresponding address: University of British Columbia, 333-6356 Agricultural Road, V6T 1Z2 Vancouver, BC, Canada.
E-mail addresses: adrover@mate.uncor.edu (J. Adrover), matias@stat.ubc.ca (M. Salibian-Barrera).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.05.005

(1) (*Stable interval.*) The minimum asymptotic coverage over the ε-contamination neighbourhood is $1-\alpha$, i.e.

$$\lim_{n\to\infty} \inf_{H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})} P_H\left(L_n < \theta < U_n\right) \geq 1-\alpha.$$

(2) (*Informative interval.*) The maximum asymptotic length of the interval is bounded over the ε-contamination neighbourhood, i.e.

$$\lim_{n\to\infty} \sup_{H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})} [U_n - L_n] < \infty.$$

It is easy to see that, for the location model, confidence intervals of the form $\overline{X}_n \pm t_{(n-1)}(\alpha/2)S_n/\sqrt{n}$ do not satisfy either Part 1 or 2 of Definition 1. The problem with the above confidence intervals is not solely due to the lack of robustness of the estimators $\overline{X}_n$ and S_n. It can be shown that even if we replace the sample mean and standard deviation by robust counterparts $\hat{\theta}_n$ and $\hat{\sigma}_n$, the resulting confidence interval only satisfies Part 2 of the above Definition.

The failure of intervals of the form $\hat{\theta}_n \pm t_{(n-1)}(\alpha/2)\hat{\sigma}_n/\sqrt{n}$ to satisfy Part 1 above is due to the fact that while the length of the interval converges to zero as $n\to\infty$, its center $\hat{\theta}_n$ may converge to a value different from the parameter of interest θ. This problem can be fixed taking into account the largest possible difference between $\hat{\theta}(H)$, the limiting value of $\hat{\theta}_n$, and the parameter of interest θ, across distributions H in the contamination neighbourhood $\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})$. This quantity is related to the maximum asymptotic bias of the estimator $\hat{\theta}_n$ (e.g. see Huber (1964)).

For the location model $Y_i = \theta + \sigma_0\,\epsilon_i$, the maximum asymptotic bias of $\hat{\theta}_n$ is

$$B(\hat{\theta}) = \sup_{H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})} \frac{\left|\hat{\theta}(H)-\theta\right|}{\sigma_0},$$

and thus, $\left|\hat{\theta}(H)-\theta\right| \leq B(\hat{\theta})\,\sigma_0$ for all $H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})$. Let $\hat{\sigma}_n$ be an estimator of σ_0 with limit $\hat{\sigma}(H)$, which in principle may be different from σ_0. For each $H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})$ we have

$$\left|\hat{\theta}(H)-\theta\right| \leq B(\hat{\theta})\sigma_0 = B(\hat{\theta})\frac{\sigma_0}{\hat{\sigma}(H)}\hat{\sigma}(H) \leq B(\hat{\theta})B_-(\hat{\sigma})\hat{\sigma}(H), \tag{3}$$

where $B_-(\hat{\sigma}) = \sup_{H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})} \sigma_0/\hat{\sigma}(H)$. Tabulated values of $B_-(\hat{\sigma})$ for different scale estimators are available in Adrover and Zamar (2004). Hence, we can estimate the largest difference $\left|\hat{\theta}(H)-\theta\right|$ using $B(\hat{\theta})B_-(\hat{\sigma})\hat{\sigma}_n$.

In the linear regression model, the maximum asymptotic bias for the slope β_1 is

$$B(\hat{\beta}_1) = \sup_{H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})} \left|\hat{\beta}_1(H)-\beta_1\right| \frac{\sigma_X}{\sigma_0},$$

where $\hat{\beta}_1(H)$ is the limit of the slope estimator $\hat{\beta}_{1,n}$ when the data have distribution H, and σ_X is the scale of the covariates under model (1). If the estimator is equivariant under affine transformations, $B(\hat{\beta}_1)$ above does not depend on the value of the parameters under the central model (see Martin et al. (1989)).

Similarly to (3), we have

$$\left|\hat{\beta}_1(H)-\beta_1\right| \leq B(\hat{\beta}_1)\frac{\sigma_0}{\sigma_X} \leq B(\hat{\beta}_1)B_-(\hat{\sigma})B_+(\hat{\sigma}_X)\frac{\hat{\sigma}(H)}{\hat{\sigma}_X(H)}, \tag{4}$$

for each $H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})$, where

$$B_-(\hat{\sigma}) = \sup_{H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})} \sigma_0/\hat{\sigma}(H) \quad \textit{and} \quad B_+(\hat{\sigma}_X) = \sup_{H\in\mathcal{H}_\varepsilon(H_{\boldsymbol{\theta}})} \hat{\sigma}_X(H)/\sigma_X,$$

and $\hat{\sigma}_X(H)$ is the limit of the scale estimator $\hat{\sigma}_{X,n}$. The quantities $B(\hat{\beta}_1)$ and $B_-(\hat{\sigma})B_+(\hat{\sigma}_X)$ on the right-hand side of (4) are available for some estimators (see Martin and Zamar (1993)) and the ratio $\hat{\sigma}(H)/\hat{\sigma}_X(H)$ can be estimated for a given sample by $\hat{\sigma}_n/\hat{\sigma}_{X,n}$.

In Adrover et al. (2004) the authors proposed robust confidence intervals of the form $\hat{\beta}_1 \pm q_n$, where q_n satisfies

$$\Phi\left(\frac{q_n-\bar{\beta}_1}{v_n}\right) + \Phi\left(\frac{q_n+\bar{\beta}_1}{v_n}\right) - 1 = 1-\alpha; \tag{5}$$

$\sqrt{n}\,v_n$ is a consistent estimator of the asymptotic variance of $\hat{\beta}_1$; and $\bar{\beta}_1$ is the estimated bias bound in (4):

$$\bar{\beta}_1 = B(\hat{\beta}_1)B_-(\hat{\sigma})B_+(\hat{\sigma}_X)\frac{\hat{\sigma}_n}{\hat{\sigma}_{X,n}}.$$

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات