Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust tests for linear regression models based on τ -estimates

Matias Salibian-Barrera^{a,*}, Stefan Van Aelst^{c,d}, Víctor J. Yohai^b

^a Department of Statistics, University of British Columbia, Vancouver, BC, Canada

^b Department of Mathematics, University of Buenos Aires, Argentina

^c Department of Mathematics, KU Leuven, Belgium

^d Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

ARTICLE INFO

Article history: Received 5 May 2014 Received in revised form 27 August 2014 Accepted 12 September 2014 Available online 20 September 2014

Keywords: Robust statistics Robust tests Linear regression

ABSTRACT

ANOVA tests are the standard tests to compare nested linear models fitted by least squares. These tests are equivalent to likelihood ratio tests, so they have high power. However, least squares estimators are very vulnerable to outliers in the data, and thus the related ANOVA type tests are also extremely sensitive to outliers. Therefore, robust estimators can be considered to obtain a robust alternative to the ANOVA tests. Regression τ -estimators combine high robustness with high efficiency which makes them suitable for robust inference beyond parameter estimation. Robust likelihood ratio type test statistics based on the τ -estimates of the error scale in the linear model are a natural alternative to the classical ANOVA tests. The higher efficiency of the τ -scale estimates compared with other robust alternatives is expected to yield tests with good power. Their null distribution can be estimated using either an asymptotic approximation or the fast and robust bootstrap. The robustness and power of the resulting robust likelihood ratio type tests for nested linear models is studied.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

An important step in regression analysis is determining which of the available explanatory variables are relevant in the proposed model. One approach is to test whether some of the regression coefficients are different from zero or not. The standard test for linear hypotheses of this type is the well-known *F*-test based on least squares estimates. It is also the likelihood ratio test when the errors are normally distributed. Unfortunately, small deviations from this assumption may seriously affect both the least squares estimates and the corresponding *F*-test, invalidating the resulting inference conclusions. Such small perturbations in the data are very common in real applications, and as the number of variables and the complexity of the models increase, they become much more difficult to detect using diagnostic methods based on non-robust estimators.

To overcome this problem robust estimators have been proposed and studied extensively in the literature. These estimators yield reliable point estimates for the model parameters even when the ideal distributional assumptions are not satisfied. Robustness properties of such estimators have been investigated via their influence function and breakdown point (see e.g. Hampel et al., 1986, Maronna et al., 2006). The influence function provides information on the effect of a

* Corresponding author. E-mail addresses: matias@stat.ubc.ca, msalibian@yahoo.ca (M. Salibian-Barrera).

http://dx.doi.org/10.1016/j.csda.2014.09.012 0167-9473/© 2014 Elsevier B.V. All rights reserved.







small amount of contamination on the estimator while, intuitively speaking, the breakdown point is the largest fraction of arbitrary contamination that can be present in the data without driving the bias of the estimator to infinity. Another robustness criterion to compare estimators is the maximum asymptotic bias which measures the effect of a positive (noninfinitesimal) fraction of contamination (Martin et al., 1989; Berrendero et al., 2007). Robust high-breakdown estimators of the regression parameters include the least median of squares and least trimmed squares estimators (Rousseeuw, 1984), S-estimators (Rousseeuw and Yohai, 1984), MM-estimators (Yohai, 1987), τ -estimators (Yohai and Zamar, 1988) and CM-estimators (Mendes and Tyler, 1996).

In this paper we consider the problem of performing inference for a linear regression model using robust estimators. Specifically, we are interested in obtaining robust and efficient tests for linear hypotheses on the regression coefficients. Robust hypothesis tests have received much less attention in the literature than point estimators. A natural approach to obtain robust tests is to use a robust point estimator of the model parameters. Robust Wald-, scores- and likelihood-ratio-type tests based on M- and GM-estimators have been proposed in the literature by Markatou and Hettmansperger (1990), Markatou et al. (1991), Markatou and He (1994), and Heritier and Ronchetti (1994). Unfortunately, the breakdown point of these estimators is less than 1/p, where p is the number of regression coefficients. In other words, the more explanatory variables in the model the less robust these estimators are. This lack of robustness in turn affects the associated test statistics.

Alternatively, one can note that the classical *F*-test compares the residual sum of squares obtained under the null and alternative hypotheses. Since the residual sum of squares can also be thought of as (non-robust) residual scale estimates, it is natural to consider test statistics of the form

$$\left(\frac{\hat{\sigma}_0^2 - \hat{\sigma}_a^2}{\hat{\sigma}_a^2}\right) \tag{1}$$

where $\hat{\sigma}_0$ is a robust residual scale estimate obtained under the null hypothesis, and $\hat{\sigma}_a$ is the scale estimate for the unrestricted model. τ -estimators (Yohai and Zamar, 1988) are a class of high-breakdown and highly efficient regression estimators that are naturally accompanied by an associated estimator of the error scale which is also highly robust and highly efficient. This is an advantage compared to other classes of high-breakdown, highly efficient regression estimators such as MM-estimators or CM-estimators. The τ -estimators are defined as the minimizers of a robust and efficient scale estimator of the regression errors. These estimators can be tuned to simultaneously have a high-breakdown point (50%) and achieve high-efficiency (e.g. 85% or 95%) at the central model with normal errors. Good robustness properties of τ -estimators have been shown for both the estimator of the regression coefficients (Berrendero and Zamar, 2001) and the estimator of the error scale (Van Aelst et al., 2013). We expect that the good robustness properties of the τ -scale estimates are compared with other scale estimators to yield tests with good robustness properties as well. Until recently, the main drawback of τ -estimators was the lack of a good algorithm for their computation. However, Salibian-Barrera et al. (2008) proposed an efficient algorithm for these estimators and implementations in R (2013) and MATLAB (2013)/OCTAVE (2014) are publicly available online at http://www.stat.ubc.ca/~matias.

In what follows we study ANOVA-type tests of the intuitively appealing test in (1) using τ -scale estimators which we call ANOVA τ -tests. We show that under certain regularity conditions the test statistics proposed in this paper are asymptotically central chi-squared distributed under the null hypothesis, and non-central chi-squared distributed under sequences of contiguous alternatives. Note that these ANOVA-type test statistics thus have a much simpler asymptotic distribution than several robust likelihood ratio type test statistics based on M-estimators whose asymptotic distribution is a linear combination of χ_1 distributions (Ronchetti, 1982; Heritier and Ronchetti, 1994). Furthermore, we derive the influence functions of these tests, which show that the tests are robust against vertical outliers and bad leverage points, although good leverage points may have a larger influence on the test statistic and corresponding level and power.

Since the finite-sample distribution of test statistics of this form is generally unknown, *p*-values are usually approximated using the asymptotic distribution of the test statistic. However, numerical experiments show that in some cases these approximations are reliable only for relatively large sample sizes. Moreover, some of the required regularity assumptions may not hold when the data contain outliers, which compromises the validity of the asymptotic approximation. To obtain better *p*-value approximations one can consider using the bootstrap (Efron, 1979). However, bootstrapping robust estimators when the data may contain outliers presents two important challenges. First, re-calculating many times the estimator is highly computationally demanding. Second, the bootstrap results may not be reliable due to bootstrap samples containing many more outliers than the original sample. In fact, this might cause the bootstrapped estimate to break down even if the original estimate did not. Salibian-Barrera and Zamar (2002) proposed a fast and robust bootstrap (FRB) method for MM-regression estimates that solves both of these problems by calculating a fast approximation to the bootstrapped MM-estimators. This FRB method has been extended to other settings (see e.g. Van Aelst and Willems, 2005, Salibian-Barrera et al., 2006 and Samanta and Welsh, 2013).

In this paper we also extend the FRB methodology to the class of τ -estimators. Salibian-Barrera (2005) showed that the FRB also works well as a way to obtain *p*-value estimates for robust scores-type test statistics. However, because the likelihood ratio type tests have a higher order of convergence than the scores-type tests, the consistency of the FRB estimator for the null distribution of the robust likelihood ratio type tests proposed here needs to be studied carefully. This problem is discussed in Section 3, where we show that the test statistics satisfy a sufficient condition given in Van Aelst and Willems (2011) that guarantees that the FRB is a consistent estimator for their null distribution. This result allows us to propose a computationally feasible and reliable *p*-value estimate based on the FRB that is consistent under weaker

دريافت فورى 🛶 متن كامل مقاله

- 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
- ISIArticles مرجع مقالات تخصصی ایران