Contents lists available at ScienceDirect

Economics Letters

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

A Hausman-type test to detect the presence of influential outliers in regression analysis

Catherine Dehon [a], Marjorie Gassner [a], Vincenzo Verardi [a,b,*]

[a] ECARES and CKE, Université Libre de Bruxelles, B-1050 Brussels, Belgium
[b] CRED, University of Namur, B-5000 Namur, Belgium

ARTICLE INFO

Article history:
Received 30 April 2008
Received in revised form 20 May 2009
Accepted 22 May 2009
Available online 30 May 2009

Keywords:
Hausman test
Outlier detection
Robustness
S-estimator

JEL classification:
C12
C13

ABSTRACT

In regression analysis, classical estimations may be excessively influenced by a few atypical observations. We propose a Hausman-type test to balance robustness and efficiency and to check whether a robust method should be implemented. An economic application is presented.

1. Introduction

In applied research, it is well known that if even a small amount of data behaves differently from the vast majority of the observations, classical estimations may be distorted, leading to results that are not representative of the population. To cope with this, several robust procedures have been proposed (see Maronna et al., 2006, for a thorough review). However, the price to pay for robustness is a loss in efficiency. It is thus not always preferable to call on robust estimators, especially if the contamination is not too severe. The goal of this paper is to provide a simple test to balance robustness and efficiency and determine whether a classical or a robust estimation procedure should be preferred. The test is basically an extension of the well known Hausman test in the context of outlier detection.

The paper is divided into five sections. After this short introduction, in Section 2 we explain the logic behind the test and in Section 3 we carry out some simulations to check its power. In Section 4 we present a simple empirical application and in Section 5, we conclude.

2. A Hausman-type test

Assume we want to estimate a regression model of the type

$$y_i = X_i\theta + \varepsilon_i \quad \text{for } i = 1,\ldots,n \tag{1}$$

where n is the sample size, X the matrix of the explanatory variables, y the dependent variable, θ the vector of regression parameters and ε_i the error term. Errors are assumed independent of the explanatory variables and i.i.d. according to the normal distribution $N(0,\sigma^2)$. Vector θ is generally estimated by ordinary least-squares (LS), the objective of which is to minimize the sum of the squared residuals. More precisely:

$$\hat{\theta}_{LS} = \underset{\hat{\theta}}{\arg\min} \sum_{i=1}^{n} r_i^2 \text{ where } r_i = y_i - X_i\hat{\theta} \tag{2}$$

Given that the square function awards excessive weight to large residuals, LS is extremely sensitive to extreme values and might lead to poor estimations if outliers are present in the dataset. To cope with this, several alternative methods have been proposed. We chose to focus on S-estimators here because of their strong robustness (largely described in the statistical literature) and good asymptotic properties.

The intuition behind this method is simple: the objective of LS is to minimize the variance of the residuals. However, since the variance is

[*] Corresponding author. CRED, University of Namur, B-5000 Namur, Belgium. Tel.: +32 0 81 725308.
E-mail address: vverardi@fundp.ac.be (V. Verardi).

0165-1765/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2009.05.017

not a robust estimator of spread, LS breaks down in the presence of outliers. The idea behind S-estimators is to minimize another measure of the dispersion of the residuals, less sensitive to extreme values. More precisely, S-estimators of regression are defined by

$$\hat{\theta}_S = \underset{\hat{\theta}}{\arg\min} \quad s\left(r_1\left(\hat{\theta}\right), \ldots, r_n\left(\hat{\theta}\right)\right) \tag{3}$$

where s is a robust measure of dispersion. The robust spread is generally estimated by an M-estimator of scale which can be seen as a robustified version of the variance. Indeed,[1] the variance of residuals is defined as $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2(\theta)$ which can be rewritten as $\frac{1}{n}\sum_{i=1}^{n} \left(\frac{r_i(\theta)}{\hat{\sigma}}\right)^2 = 1$. LS thus aims at finding the minimum value of $\hat{\sigma}$ that satisfies the latter equation. Again, the square may distort things as it awards considerable importance to large residuals. To increase robustness, the square function could be replaced by some other function ρ that awards less importance to large residuals.[2] The estimation problem then consists in finding the smallest $\hat{\sigma}$ (that we call $\hat{\sigma}^S$ to avoid any confusion) that satisfies $\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}^S}\right) = 1$. Obviously, if the data are Gaussian, this estimator should coincide with the standard deviation. A correction factor b is thus needed to guarantee this consistency. The problem now boils down to finding the minimal $\hat{\sigma}^S$ that satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\theta)}{\hat{\sigma}^S}\right) = b \tag{4}$$

where b is set at $E_\Phi[\rho]$ (where Φ is the standard Normal cumulative function).

In this paper, the ρ function considered is Tukey's Biweight

$$\rho(r) = \begin{cases} \dfrac{1}{6c^4}r^6 - \dfrac{1}{2c^2}r^4 + \dfrac{1}{2}r^2 & \text{if } |r| \le c \\ \dfrac{c^2}{6} & \text{if } |r| > c \end{cases} \tag{5}$$

where the tuning parameter c is set at 1.547 in order to ensure resistance to a contamination of up to 50% of the observations (it is then said to have a breakdown point of 50%). The price to pay for such high resistance to outliers is very low Gaussian efficiency (28.7%).

To summarize, the LS-estimator is efficient but not robust while the S-estimator is very robust but inefficient.

An interesting question is then how to decide in each situation which method is preferable. A Hausman-type test may be helpful in this context. Indeed, assuming that Gauss–Markov hypotheses are respected, it will allow to test if the gain in consistency is not dampened by an excessive loss in efficiency.

The original Hausman test (1978) is based on comparing an estimator which is efficient under H_0 but inconsistent under H_1, with an estimator that is consistent both under H_0 and H_1, but inefficient. Even though the problem is not the same here, we retain the same underlying logic since we compare the classical LS-estimator ($\hat{\theta}_{LS}$) which is consistent and efficient under the null of no inconsistency due to outliers (but inconsistent under the alternative) to the S-estimator ($\hat{\theta}_S$) which is consistent under both H_0 and H_1, but inefficient.

The probability limit of the difference between the two estimators (defined as $\hat{q}$) is zero if and only if no outlier is present. As far as the variance of $\hat{q}$ is concerned, Hausman (1978) proved that when two estimators (one which is consistent but inefficient, the other efficient but not necessarily consistent) are correlated, the asymptotic variance of their difference is given by the difference of their respective variances.

Since it is well known that (under H_0)

$$\hat{\theta}_{LS} \overset{a}{\sim} N\left(\theta, \sigma^2 (X'X)^{-1}\right) \tag{6}$$

and since Rousseeuw and Yohai (1984) proved that

$$\hat{\theta}_S \overset{a}{\sim} N\left(\theta, \frac{\sigma^2 (X'X)^{-1}}{e}\right) \tag{7}$$

where e is the efficiency of the S-estimator, the asymptotic variance of $\hat{q}$, denoted by $V(\hat{q})$, can be written as

$$V(\hat{q}) = V\left(\hat{\theta}_S\right) - V\left(\hat{\theta}_{LS}\right) = \frac{\sigma^2 (X'X)^{-1}}{e} - \sigma^2 (X'X)^{-1}. \tag{8}$$

The nuisance parameter σ is not known. We therefore estimate it with $\hat{\sigma}^S$ (estimated on robust residuals). In this way, we obtain a consistent estimator of $V(\hat{q})$ that we call $\hat{V}(\hat{q})$.

The Hausman test statistic is now obtained by

$$H = \hat{q}' \left[\hat{V}(\hat{q})\right]^{-1} \hat{q} \tag{9}$$

Hausman (1978) shows that under the null, H is asymptotically distributed as a central χ^2_p where p is the number of unknown regression parameters.[3] If the latter statistic is higher than the tabulated value of a χ^2_p at a given level of confidence, we reject the hypothesis that the difference between the estimators is not systematic and thus reject the LS estimator. Otherwise, we conclude that the efficiency loss resulting from the use of the S-estimator is more costly than the bias produced by the use of LS and retain the latter. We implemented this test in Stata, R and Matlab. It is available from the authors upon request.

3. Simulations

This section studies the behavior of the test under contamination (H_1). In linear regressions, outliers are classified into three categories: *bad leverage points*, *good leverage points* and *vertical outliers* (see Fig. 1(a)). We study the power of the test under these three types of contamination. Note that, in the context of a least-squares regression, bad leverage points influence the estimation of all coefficients heavily, vertical outliers have a significant influence on the estimation of the intercept but their effect on slope coefficients is rather mild and the effect of good leverage points is marginal on all coefficients (some authors even state that they tend to stabilize the regression hyperplane).

For the simulations, observations were generated according to the model

$$y_i = \theta_0 + x_i + \varepsilon_i \tag{10}$$

where $x \sim N(0,1)$, $\varepsilon \sim N(0,1)$ and $\theta_0 = 1$. The chosen sample sizes are $n = 100$, 200, 500 and 700. For all simulations under the alternative, we introduce a very small percentage of contamination, 1%. If the

[1] Under Gauss–Markov assumptions.

[2] ρ must be even, non decreasing for positive values, less increasing than the square with a unique minimum at zero.

[3] Simulations tend to show that this asymptotic result works reasonably well for relatively small sample sizes of the order of $n = 200$.