Extending the Hausman Test to Check for the presence of Outliers

Catherine Dehon,
ECARES, SBS-EM, Université Libre de Bruxelles

Marjorie Gassner
ECARES, SBS-EM and CKE Université Libre de Bruxelles

Vincezo Verardi
CRED, Facultés Universitaires Notre-Dame de la Paix, Namur
and ECARES and CKE, Université Libre de Bruxelles

ECARES working paper 2011-036

ECARES
ULB - CP 114/04
50, F.D. Roosevelt Ave., B-1050 Brussels BELGIUM
www.ecares.org

Extending the Hausman test to check for the presence of outliers*

Catherine Dehon†, Marjorie Gassner‡and Vincenzo Verardi§

November 10, 2011

Abstract

In this paper, we follow the same logic as in Hausman (1978) to create a testing procedure that checks for the presence of outliers by comparing a regression estimator that is robust to outliers (S-estimator), with another that is more efficient but affected by them. Some simulations are presented to illustrate the good behavior of the test for both its size and its power.

KEYWORDS: S-estimators, MM-estimators, Outliers, Linear Regression, Generalized Method of Moments, Robustness.
JEL CLASSIFICATION: C12, C21, H11

1 Introduction

In a seminal paper, Hausman (1978) introduced a testing procedure that, under some assumptions, allows to balance consistency and efficiency when comparing two estimators. Hausman's testing procedure is used extensively in econometrics: in the context of panel data, for example, it is called on to check whether the assumptions underlying the random-effects model are satisfied. This is done by comparing the fixed-effects model (consistent, but inefficient) with the random-effects model (more efficient, but potentially inconsistent if a set of assumptions is not fulfilled). If the differences between the corresponding coefficients of the two models is not systematic, the test indicates that it is preferable to use a random-effects model since the gain in efficency dominates the loss in consistency. In

*The authors would like to thank Christophe Croux for his insightful comments. Catherine Dehon gratefully acknowledges research support from FRFC (Fonds de Recherche Fondamentale Collective) and from the ARC contract of the Communauté Française de Belgique. Catherine Dehon is also member of ECORE, the association between CORE and ECARES. Vincenzo Verardi is Associated Researcher of the FNRS and gratefully acknowledges their financial support.

†Université libre de Bruxelles, ECARES, SBS-EM, 50, avenue F. Roosevelt, CP 114/04, 1050 Brussels Tel.: +32-2-6503858 Fax: +32-2-6504012

‡Université libre de Bruxelles, ECARES and CKE, SBS-EM, 50, avenue F. Roosevelt, CP 139, 1050 Brussels. Tel.: +32-2-6503843. Email: mgassner@ulb.ac.be

§University of Namur (CRED) and Université libre de Bruxelles (ECARES and CKE). Rempart de la Vierge, 8. B-5000 Namur. E-mail: vverardi@fundp.ac.be.

this paper, we follow the same logic as in Hausman (1978) to create a testing procedure that enables to check if the presence of outliers influences the estimation of the regression parameters in a linear model. The idea is to compare a regression estimator that is robust (S), with an estimator that has higher efficiency but is more influenced by outliers (hereafter called MM, not to be mistaken for the exactly identified Generalized Method of Moments estimator that will be denoted by GMM). More precisely, consider the regression model

$$Y_i = X_i^t \theta + \varepsilon_i$$

where Y_i is the dependent variable and X_i is the $((p+1) \times 1)$ vector of covariates (plus the constant) observed for $i = 1, ...N$. The testing procedure consists in comparing the regression coefficients respectively estimated by the S- and MM-estimators to check if they are statistically different (as will be explained later, the constant is disregarded). The above-mentioned comparison of the regression coefficients is carried out by calling on the Generalized Hausman test statistic defined as

$$H = (\hat{\theta}^S - \hat{\theta}^{MM})^t [Var(\hat{\theta}^S - \hat{\theta}^{MM})]^{-1} (\hat{\theta}^S - \hat{\theta}^{MM}) \quad (1)$$

where $\hat{\theta}^S$ and $\hat{\theta}^{MM}$ represent respectively the S- and MM-estimators of θ (with a given Gaussian efficiency). Since the Generalized Hausman statistic is asymptotically distributed as a χ^2_p, where p is the number of covariates, it is possible to set an upper bound above which the estimated parameters can be considered as statistically different: if the value of H is above $\chi^2_{p,(1-\alpha)}$ (where α is the given significance level), the difference between $\hat{\theta}^S$ and $\hat{\theta}^{MM}$ (and hence the lack of robustness of MM) is too large with respect to the gain in efficiency.

For this testing procedure to be operational, we need an estimate of the variance of the difference $(\hat{\theta}^S - \hat{\theta}^{MM})$ that remains consistent under heteroskedasticity and/or asymmetry. This paper aims at developing a modified Hausman testing procedure allowing not only to compare S-estimators with MM-estimators (with a given efficiency level), but also to detect the presence of outliers by comparing

S-estimators with non-robust LS-estimators (a limit case of MM). The structure of the paper is the following: after the first introductory section, in Section 2 we develop the robustness test. In Section 3 we run some simulations to observe its behavior in finite samples and in Section 4 we conclude.

2 General testing procedure

Consider the regression model

$$Y_i = X_i^t\theta + \varepsilon_i$$

where Y_i is the dependent variable, X_i is the $((p+1)\times 1)$ vector of covariates observed for $i = 1, ..., N$ and σ is the dispersion of ε. To estimate parameter column vector θ, a measure s of the dispersion of the residuals $r_i(\theta) = Y_i - X_i^t\theta$ for $1 \leq i \leq n$ is minimized. The regression estimate $\hat{\theta}_0$ can then be defined by

$$\hat{\theta}_0 = \arg\min_{\theta} \ s(r_1(\theta), ..., r_n(\theta)). \tag{2}$$

In the case of LS, the measure of dispersion that is minimized is the (squared root of the) variance. The problem with LS is that an excessive importance is awarded to observations with very large residuals and, consequently, the estimated parameters are distorted if outliers are present. To take this into account, Rousseeuw and Yohai (1984) propose to minimize another measure of dispersion s of the residuals, an M-estimator of scale (s), defined as the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0(\frac{r_i(\theta)}{s}) = \delta \tag{3}$$

where $\delta = E[\rho_0(Z)]$ with $Z \sim N(0,1)$ where $\rho_0(\cdot)$ function is even, non decreasing for positive values and less increasing than the square. This is equivalent to solving

$$\begin{cases} \min\limits_{\theta} s(r_1(\theta), ..., r_n(\theta)) \\ \text{s.t.} \frac{1}{n} \sum\limits_{i=1}^{n} \rho_0 \left(\frac{Y_i - X_i^t \theta}{s} \right) = \delta \end{cases} \tag{4}$$

yielding solutions $\hat{\theta}_0$ and $\hat{\sigma}$ such that

$$\begin{cases} \frac{1}{n} \sum\limits_{i=1}^{n} \rho_0' \left(\frac{Y_i - X_i^t \hat{\theta}_0}{\hat{\sigma}} \right) X_i^t = 0 \\ \frac{1}{n} \sum\limits_{i=1}^{n} \rho_0 \left(\frac{Y_i - X_i^t \hat{\theta}_0}{\hat{\sigma}} \right) = \delta \end{cases} \tag{5}$$

where ρ_0' is the first derivative of ρ_0. If ρ_0 is the square function (and $\delta = 1$), this becomes a standard LS maximization problem.

The choice of $\rho_0(\cdot)$ is crucial to guarantee robustness and high Gaussian efficiency. The function ρ_0 usually used in (3) is the Tukey Biweight function defined as

$$\rho_0(u) = \begin{cases} \frac{k^2}{6} \left(1 - \left[1 - \left(\frac{u}{k} \right)^2 \right]^3 \right) \text{ if } |u| \leq k \\ \frac{k^2}{6} \text{ if } |u| > k \end{cases} . \tag{6}$$

If the tuning parameter k is set at 1.547, it can be shown that the breakdown point (i.e. the maximal contamination an estimator can withstand before breaking) reaches 50%. The Gaussian efficiency is however rather low (28%). To increase the efficiency, Rousseeuw and Yohai (1984) and Yohai (1987) introduced MM-estimators that combine a high-breakdown point and high efficiency. These estimates result from minimizing a loss function of the residuals $\sum_{i=1}^{n} \rho(\frac{r_i(\theta)}{\hat{\sigma}})$ where parameter σ is set at the value estimated by the S-estimator ($\hat{\sigma}$) and, as $\rho_0(\cdot)$, the function $\rho(\cdot)$ is even, non decreasing for positive values and less increasing than the square with $\rho(\cdot) \leq \rho_0(\cdot)$. The estimate $\hat{\theta}$ is defined by:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \rho(\frac{r_i(\theta)}{\hat{\sigma}}).$$

Values $\hat{\theta}$, $\hat{\theta}_0$ and $\hat{\sigma}$ are such that

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{Y_i-X_i^t\hat{\theta}}{\hat{\sigma}}\right)X_i^t=0 \\ \frac{1}{n}\sum_{i=1}^{n}\rho_0'\left(\frac{Y_i-X_i^t\hat{\theta}_0}{\hat{\sigma}}\right)X_i^t=0 \\ \frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{Y_i-X_i^t\hat{\theta}_0}{\hat{\sigma}}\right)=\delta \end{cases} \tag{7}$$

where ψ is ρ', the first derivative of ρ.

It is common to also use a Tukey Biweight $\rho(\cdot)$ function for the final MM-estimator where the tuning constant can be modified to attain a Gaussian efficiency much higher than 28%. For example, if $k = 4.685$, the Gaussian efficiency is 95% and if $k = 6.256$ it is 99%. For the sake of clarity, we denote by ρ_0 (and ρ_0') the Tukey Biweight function (and its first derivative) in which the tuning parameter is set to 1.547, the function used for the preliminary S-estimator. On the other hand, we use the general notation of ρ (and ψ) for the Tukey Biweight function (and its first derivative) used in the final estimator of the MM where tuning parameter is set according to the desired Gaussian efficiency. It might thus be tempting to only consider highly efficient MM-estimators. This is not advised since the associated bias might be large even if the estimator does not break (see Maronna et al. 2006). As a consequence, it is of the utmost importance to find the highest efficiency without paying the price of an excessive bias. The test we propose hereunder can be used to achieve this as it allows to determine which MM-estimators are statistically different from S (and hence excessively biased).

From (7) and as shown by Croux et al. (2003), MM-estimators are first-order equivalent with exactly identified Generalized Method of Moments estimators (GMM) for $\vartheta = (\theta^t, \theta_0^t, \sigma)^t$ with moment function m_i (for observation i)

$$m_i(\vartheta) = \begin{pmatrix} \psi(\varepsilon_i)X_i \\ \rho_0'(\varepsilon_{0i})X_i \\ \rho_0(\varepsilon_{0i})-\delta \end{pmatrix}, \text{ from here on abbreviated by } \begin{pmatrix} \psi_i X_i \\ \rho_{0i}' X_i \\ \rho_{0i}-\delta \end{pmatrix}$$

where $\varepsilon_i = \frac{Y_i - X_i^t\theta}{\sigma}$ and $\varepsilon_{0i} = \frac{Y_i - X_i^t\theta_0}{\sigma}$. To clarify notations, we chose to denote by θ_0 the regression parameter that is estimated by the S-estimator and by θ, the parameter estimated by MM.

Following Hansen (1982), Croux et al. (2008) show that $\hat{\vartheta}$ has a limiting normal distribution given by

$$\sqrt{N}(\hat{\vartheta} - \vartheta) \longrightarrow N(0, V)$$

where, defining $G = E\left[\frac{\partial m_i(\boldsymbol{\vartheta})}{\partial \vartheta^t}\right]$ and $\Omega = E[m_i(\vartheta)m_i^t(\vartheta)]$, the asymptotic variance V is

$$V = G^{-1}\Omega(G^t)^{-1} \tag{8}$$

Since $\Omega = E\begin{pmatrix} \psi_i^2 X_i X_i^t & \psi_i \rho'_{0i} X_i X_i^t & \psi_i \rho_{0i} X_i \\ \rho'_{0i}\psi_i X_i X_i^t & (\rho'_{0i})^2 X_i X_i^t & \rho'_{0i}\rho_{0i} X_i \\ \rho_{0i}\psi_i X_i^t & \rho_{0i}\rho'_{0i} X_i^t & (\rho_{0i})^2 - \delta^2 \end{pmatrix}$

and $G^{-1} = -\begin{pmatrix} \sigma[E(\psi'_i X_i X_i^t)]^{-1} & 0 & -\sigma[E(\psi'_i X_i X_i^t)]^{-1}E(\psi'_i X_i \varepsilon_i)[E(\rho'_{0i}\varepsilon_{0i})]^{-1} \\ 0 & \sigma[E(\rho''_{0i} X_i X_i^t)]^{-1} & -\sigma[E(\rho''_{0i} X_i X_i^t)]^{-1}E(\rho''_{0i} X_i \varepsilon_{0i})[E(\rho'_{0i}\varepsilon_{0i})]^{-1} \\ 0 & 0 & \sigma[E(\rho'_{0i}\varepsilon_{0i})]^{-1} \end{pmatrix}$

defining

$A = \sigma[E(\psi'_i X_i X_i^t)]^{-1}$;

$a = AE(\psi'_i X_i \varepsilon_i)[E(\rho'_{0i}\varepsilon_{0i})]^{-1}$;

$B = \sigma[E(\rho''_{0i} X_i X_i^t)]^{-1}$ and

$b = BE(\rho''_{0i} X_i \varepsilon_{0i})[E(\rho'_{0i}\varepsilon_{0i})]^{-1}$,

(8) yields the asymptotic variances and covariances i.e.

$$Var(\hat{\theta}^{MM}) = AE(\psi_i^2 X_i X_i^t)A - aE(\psi_i \rho_{0i} X_i^t)A - AE(\psi_i X_i \rho_{0i})a^t + aE((\rho_{0i})^2 - b^2)a^t$$

$$Var(\hat{\theta}^{S}) = BE((\rho'_{0i})^2 X_i X_i^t)B - bE(\rho'_{0i}\rho_{0i} X_i^t)B - BE(\rho'_{0i}\rho_{0i} X_i)b^t + bE((\rho_{0i})^2 - b^2)b^t$$

$$Cov(\hat{\theta}^{S}, \hat{\theta}^{MM}) = AE(\psi_i \rho'_{0i} X_i X_i^t)B - aE(\rho'_{0i}\rho_{0i} X_i^t)B - AE(\psi_i X_i \rho_{0i})b^t + aE((\rho_{0i})^2 - b^2)b^t.$$

Estimating the (co)variances by $\widehat{Var}(\hat{\theta}^{MM})$, $\widehat{Var}(\hat{\theta}^{S})$ and $\widehat{Cov}(\hat{\theta}^{S}, \hat{\theta}^{MM})$, it is straightforward to compare the S-estimator with the MM-estimator by using the Generalized Hausman statistic defined by (1) with $Var(\hat{\theta}^{S} - \hat{\theta}^{MM}) = Var(\hat{\theta}^{S}) + Var(\hat{\theta}^{MM}) - 2Cov(\hat{\theta}^{S}, \hat{\theta}^{MM})$ i.e.

$$H = (\hat{\theta}^{S} - \hat{\theta}^{MM})^t[\widehat{Var}(\hat{\theta}^{S}) + \widehat{Var}(\hat{\theta}^{MM}) - 2\widehat{Cov}(\hat{\theta}^{S}, \hat{\theta}^{MM})]^{-1}(\hat{\theta}^{S} - \hat{\theta}^{MM}) \tag{9}$$

In this way, we test the null hypothesis that an MM-estimator with a given level of efficiency is not statistically different from an S-estimator and hence should be preferred due to its higher efficiency.

Since Gervini and Yohai (2002) showed that, in the presence of outliers, only slopes can be satisfactorily estimated when the error distribution is asymmetric, the test will be based on the comparison of the slope estimated parameters and the constant will be disregarded.

2.1 Outlier identification test

Since LS is the special case of the MM-estimator, where, in the corresponding Tukey biweight function, $k \longrightarrow \infty$, $\rho(\varepsilon) = \frac{\varepsilon^2}{2}$, thus $\psi(\varepsilon) = \varepsilon$ and $\psi'(\varepsilon) = 1$, equation (9) can be directly used to test if outliers have distorted classical regression parameters. The values of A and a become $A = \sigma[E(X_i X_i^t)]^{-1}$ and $a = AE(X_i\varepsilon_i)[E(\rho'_{0i}\varepsilon_{0i})]^{-1}$ while those of B and b remain unchanged. As a consequence,

$$Var(\hat{\theta}^{LS}) = AE(\varepsilon_i^2 X_i X_i^t)A - aE(\varepsilon_i\rho_{0i}X_i^t)A - AE(\varepsilon_i X_i\rho_{0i})a^t + aE((\rho_{0i})^2 - b^2)a^t$$

$$Cov(\hat{\theta}^{S}, \hat{\theta}^{LS}) = AE(\varepsilon_i\rho'_{0i}X_i X_i^t)B - aE(\rho'_{0i}\rho_{0i}X_i^t)B - AE(\varepsilon_i X_i\rho_{0i})b^t + aE((\rho_{0i})^2 - b^2)b^t$$

while $Var(\hat{\theta}^{S})$ remains unchanged.

By replacing $\widehat{Var}(\hat{\theta}^{MM})$ by $\widehat{Var}(\hat{\theta}^{LS})$ and $\widehat{Cov}(\hat{\theta}^{S}, \hat{\theta}^{MM})$ by $\widehat{Cov}(\hat{\theta}^{S}, \hat{\theta}^{LS})$ in (9) we can check whether the difference between the coefficients in the S- and LS-estimators is systematic or not. If the null is rejected, the influence of the outliers is such that the gained efficiency associated with a classical estimator is not sufficient to balance the corresponding bias (due to outliers). In such a case, a robust estimator should be preferred. On the other hand, if it is not rejected, the influence of the outliers is clearly rather limited, implying that a classical estimator will be only mildly biased and should be preferred to a robust one given its higher statistical precision.

In the particular case of symmetric errors and homoskedasticity, this test simplifies to the test proposed by Dehon et al. (2009a) i.e.

Proposition 1 *If the error term is symmetric and homoskedastic and* $k \longrightarrow \infty$*, then* $Cov\left(\hat{\theta}^{S}, \hat{\theta}^{LS}\right) = Var\left(\hat{\theta}^{LS}\right)$.

Proof. When $k \longrightarrow \infty$, $\rho(\varepsilon) = \frac{\varepsilon^2}{2}$, thus $\psi(\varepsilon) = \varepsilon$ and $\psi'(\varepsilon) = 1$.

From the symmetry and homoskedasticity hypotheses, $a = b = 0$, thus $A = \sigma[E(\psi'_i X_i X_i^t)]^{-1} = \sigma[X_i X_i^t]^{-1}$and $B = \sigma[E(\rho''_{0i}X_i X_i^t)]^{-1} = \sigma E(\rho''_{0i})^{-1}[X_i X_i^t]^{-1}$ and

$$\begin{aligned} Cov(\hat{\theta}^{S}, \hat{\theta}^{MM}) &= AE(\psi_i\rho'_{0i}X_i X_i^t)B = \sigma[X_i X_i^t]^{-1}E(\psi_i\rho'_{0i}X_i X_i^t)\sigma E(\rho''_{0i})^{-1}[X_i X_i^t]^{-1} \\ &= \sigma[X_i X_i^t]^{-1}E\left(\varepsilon\left(\tfrac{\varepsilon^5}{k^4} - \tfrac{2\varepsilon^3}{k^2} + \varepsilon\right)\right)X_i X_i^t.\sigma\left\{E\left(\tfrac{5\varepsilon^4}{k^4} - \tfrac{6\varepsilon^2}{k^2} + 1\right)\right\}^{-1}[X_i X_i^t]^{-1} \\ &= \sigma^2[X_i X_i^t]^{-1}\left(\tfrac{1}{k^4}E\left(\varepsilon^6\right) - \tfrac{2}{k^2}E\left(\varepsilon^4\right) + E\left(\varepsilon^2\right)\right)X_i X_i^t.\left(\tfrac{5}{k^4}E\left(\varepsilon^4\right) - \tfrac{6}{k^2}E\left(\varepsilon^2\right) + 1\right)^{-1}[X_i X_i^t]^{-1} \end{aligned}$$

$$= \sigma^2[X_i X_i^t]^{-1} \left(\tfrac{15}{k^4} - \tfrac{6}{k^2} + 1\right) . \left(\tfrac{15}{k^4} - \tfrac{6}{k^2} + 1\right)^{-1} X_i X_i^t [X_i X_i^t]^{-1}$$
$$= \sigma^2[X_i X_i^t]^{-1} = Var\left(\hat{\theta}^{LS}\right) . \blacksquare$$

From there, $H = (\hat{\theta}^S - \hat{\theta}^{LS})^t [Var\left(\hat{\theta}^S\right) - Var\left(\hat{\theta}^{LS}\right)]^{-1} (\hat{\theta}^S - \hat{\theta}^{LS})$ which is the test statistic proposed by Dehon et al. (2009a).

In the following section, we run some simulations to check how the test behaves in finite samples. Before that, we briefly present the a robust alternative test that is available in the literature (see Yohai et al., 1991). It will serve as the comparison benchmark for the simulations. The test proposed by Dehon et al. (2009a) is not considered in the simulation since it is nothing else than a specific case of the one we propose here.

3 Simulations

3.1 The Yohai, Stahel and Zamar (1991) test

In 1991, Yohai, Stahel and Zamar developed a test (YSZ) to compare the behavior of an S-estimator with that of an MM-estimator (with a given efficiency), based on the scale of the residuals. The test statistic they propose is

$$T = \frac{2n(\hat{\sigma}^S_{MM} - \hat{\sigma}^S)}{\upsilon_0 d^2 (\hat{\sigma}^S)^2} \tag{10}$$

where n is the number of observations, $\hat{\sigma}^S_{MM}$ is the M-estimator of scale of the residuals (defined in eq. 3) fitted by the MM-estimator, $\hat{\sigma}^S$ is the M-estimator of scale of the residuals fitted by the S-estimator, $\tilde{r}_i$ are the robust standardized residuals fitted by the S-estimator, $\upsilon_0 = \frac{\Sigma \rho_0''(\tilde{r}_i)}{\hat{\sigma}^S_{MM} \Sigma \rho_0'(\tilde{r}_i)\tilde{r}_i}$ and $d^2 = \frac{1}{n}\Sigma(\frac{\rho'(\tilde{r}_i)}{(1/n)\Sigma\rho''(\tilde{r}_i)} - \frac{\rho_0'(\tilde{r}_i)}{(1/n)\Sigma\rho_0''(\tilde{r}_i)})^2$.

Using standard asymptotic theory, they show that T is asymptotically distributed as a χ^2_{p+1}. However, examining (10) two drawbacks of the test emerge: first, the test focuses on the bias of the MM-estimator. Second, it is based on the assumption of a single scale of the residuals and is thus not appropriate in case of heteroskedasticity and/or asymmetry in the error term. This test will serve as the benchmark in the simulations since it is the one commonly used to test whether an MM-estimator (with a given level of efficiency) can be safely used.

3.2 Size and power of the test

In this section, we consider two aspects of the behavior of the test we propose.

First, we study its finite-sample behavior (under the null hypothesis of no outlier contamination) by comparing: i) an MM- to an S-estimator and ii) an LS- to an S-estimator. The loss function (ρ_0) used

to compute the S-estimator (with a breakdown point of 50%) and the MM-estimator with a Gaussian efficiency set to 95% where ρ is Tukey's biweight function given in (6) with the tuning parameter set respectively to $k = 1.546$ and $k = 4.685$. We check the size of the test under three assumptions on the error term: i) homoskedastic normality, ii) heteroskedastic normality and iii) homoskedastic asymmetry.

Second, we investigate the behavior of the test under contamination. The power is computed considering the most influential type of outliers (i.e. bad leverage points).

For the size of the test we simulate the data under three different sampling schemes for the error terms (homoskedastic normality, heteroskedastic normality and asymmetry) and three different sample sizes ($n = 500$, $n = 1000$ and $n = 2000$).

More precisely, the data generating process is

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \varepsilon_i \tag{11}$$

for $i = 1, \ldots, n$. The regression parameters $\theta_0, \theta_1, \theta_2$ are set to 1. The explanatory variables x_1 and x_2 are generated as *i.i.d.* standard normal random variables. The error term ε is generated according to three different designs:

i Homoskedastic Normal errors : ε_i is generated from a standard normal distribution for $i = 1, \ldots, n$;

ii Heteroskedastic Normal errors: $\varepsilon_i = |x_{i1}|\, u_i$ for $i = 1, \ldots, n$ where u is generated from a standard normal distribution;

iii Homoskedastic Asymmetric errors: ε_i is generated from a log-normal with mean zero for $i = 1, \ldots, n$.

For each case, we generate $m = 5000$ samples of n observations.

The size of the test we propose and of that of Yohai et al. (1991), are reported in Table 2. These sizes are measured by counting the percentage of times (over repeated samples) that the test statistic is larger than a given percentile (95^{th} in our case) of a χ^2 distribution with respectively p and $(p + 1)$ degrees of freedom. Ideally they should therefore be close to 5%. The QQ-plots comparing empirical and theoretical quantiles of the χ^2 in each situation can be found at "http://homepages.ulb.ac.be/~vverardi/graphs/QQplots.pdf". From Table 2, it is clear that under Gaussian and asymmetric assumptions for the error term, the empirical level of the two versions of our test (LS versus S and MM versus S) is very close to the theoretical value of 5%. The same conclusion holds for the Yohai, Stahel and Zamar test under the assumption of normality but not in the case

of asymmetry. The situation is not as good under the specification of heteroskedastic errors that we used in the simulations since the level of our test is adequate for the comparison between the LS- and S-estimators but is slightly higher than 5% for the comparison with the MM-estimator. However, the YSZ test yields an even higher difference between empirical and theoretical levels. These results also show that since the test is asymptotic, its behavior improves when the sample size increases.

[INSERT TABLE 1 HERE]

The second part of this section is devoted to the study of the power of the test under contamination. It is well-known that points outlying in the x-dimension (design space) and that lie far away from the regression line, called leverage points, are the most "dangerous" outliers (see Dehon et al., 2009b). We therefore focus on this type of outliers in the simulations. With other types of outliers the test we propose behaves even better but its difference with respect to the benchmark becomes smaller.

For the simulations, observations were generated according to model

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i \tag{12}$$

for $i = 1, \ldots, n$ where both parameters are equal to one. The sample sizes used are $n = 500, 1000$ and 2000 and the specifications for the error term are the same as for previous simulations. For all of the replications we introduced a small percentage of contamination (5%). To generate the contamination, we replaced 5% of the x-values by an integer constant that increases in succession from 0 to 9. The biases of the LS-, MM- and S-slope estimators are computed for all types of bad leverage outliers and presented in Figure 1 (where $n = 1000$ and the error term is assumed symmetric and homoskedastic). On the left panel, the bias of LS (dotted line) is compared to that of S (dashed line) while on the right one it is the bias of MM (dotted line) that is compared to that of S (dashed line). The percentage of rejection of the null is represented by the plain line.

[INSERT FIGURE 1 HERE]

The left panel of Figure 1 shows that the bias of the LS-estimator increases rapidly when the leverage effect becomes substantial (i.e. for x-coordinates ranging from 2 to 9). On the other hand, the bias of the S-estimator remains very small, which is not surprising as the S-estimator is very robust. The percentage of rejection of the null of no contamination increases quickly to reach 100% for an x-coordinate of 3. Though we only present the homoskedastic case here, whatever the scenario (normality, heteroskedasticity or asymmetry) the test behaves comparably well. When the x-coordinate of the contamination is smaller than 1, the percentage of rejection (hence the size of the test) is close to 5%. The right panel of Figure 1 shows that the bias of the MM-estimator starts increasing proportionnally to the leverage effect. However, from a certain point on, it decreases. The reason for this is that the

MM is a redescending estimator: the importance awarded to residuals increases up to a point and then starts decreasing toward zero. The influence of outliers is therefore significant only if they are located in the neighborhood of this point which is at 4 in this case.
To get a clearer idea of the power of the test, we generated 1000 samples for each type of contamination, and for each of them computed the percentage of rejection of the null. Results are presented in Figure 2. On the left panel, the test compares S to LS, while on the right it compares S to MM.

[INSERT FIGURE 2 HERE]

The percentage of rejection for two different sample sizes ($n = 500$ and $n = 2000$) and the three scenarios for the error term are plotted in Figure 2. As expected, the test obtained by comparing the LS- and S-estimators rejects the null hypothesis more rapidly when the sample size is larger (for all scenarios). The heteroskedastic case seems to yield the least powerful result.

Concerning the comparison between the MM- and the S-estimators, again the null hypothesis is more rapidly rejected when the sample size is larger. The test behaves very well under normality or asymmetry, but is seems that the detection of outliers is more difficult with heteroskedastic errors (see Table 3).

[INSERT TABLE 2 HERE]

[INSERT TABLE 3 HERE]

4 Conclusion

The objective of the paper is to extend Hausman's (1978) specification test to outlier detection. More precisely, we adopt a similar approach to compare an estimator (S) that withstands outlier contamination (and is rather inefficient) with a more efficient but potentially inconsistent one (MM). We believe that the tradeoff between consistency and efficiency will enable to make an informed decision as to which estimator should be preferred. From a practical point of view, what we suggest is to start by testing if regression coefficients estimated by least squares (a limit case of MM), have not been excessively influenced by the presence of outliers. If they have not, least squares is the preferable method. Otherwise, we suggest to compare an S with several MMs with different efficiencies. The estimator that will ultimately be retained is the one that, while not rejecting the null, has the highest efficiency.

References

[1] Croux, C., Dhaene, G. and Hoorelbeke, D., 2003, Robust Standard Errors for Robust Estimators, Center for Economic Studies, KULeuven. Discussions Paper Series (DPS) 03.16

[2] Dehon, C., Gassner, M. and Verardi, V., 2009a, A New Hausman Type Test to Detect the Presence of Influential Outliers, Economics Letters, 105, 64-67.

[3] Dehon, C., Gassner, M. and Verardi, V., 2009b, Beware of Good Outliers and Overoptimistic Conclusions, Oxford Bulletin of Economics and Statistics, 71, 437-452.

[4] Gervini, D. and Yohai, V.J., 2002, A class of robust and fully efficient regression estimators. Annals of Statistics 30 pp. 583-616.

[5] Hansen, L.P., 1982, Large Sample Properties of Generalized Method of Moments Estimators. Econometrica, 50(4), pp. 1029-1054.

[6] Hausman, J.A., 1978, Specification Tests in Econometrics, Econometrica 46 (6), pp. 1251-1271.

[7] Maronna, R., Martin, D. and Yohai, V.J., 2006, Robust Statistics: Theory and Methods. John Wiley & Sons Ltd, England

[8] Rousseeuw, P.J. and V.J. Yohai, 1984, Robust regression by means of S-estimators. In: Franke, J., Härdle W., Martin, R.D. (Eds.), Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics 26, New York: Springer Verlag.

[9] Yohai, V.J., Stahel, W., and Zamar, R. H., 1991, A procedure for robust estimation and inference in regression. In Directions in Robust Statistics and Diagnosis, Part II, Werner Stahel and Sandford Weisberg editors, IMA volumes in Mathematics and its Applications vol. 34, pp. 365-374.

[10] Yohai, V.J., 1987, High breakdown-point and high efficiency M-estimates for regression. The Annals of Statistics, vol. 15, pp. 642-656, 1987

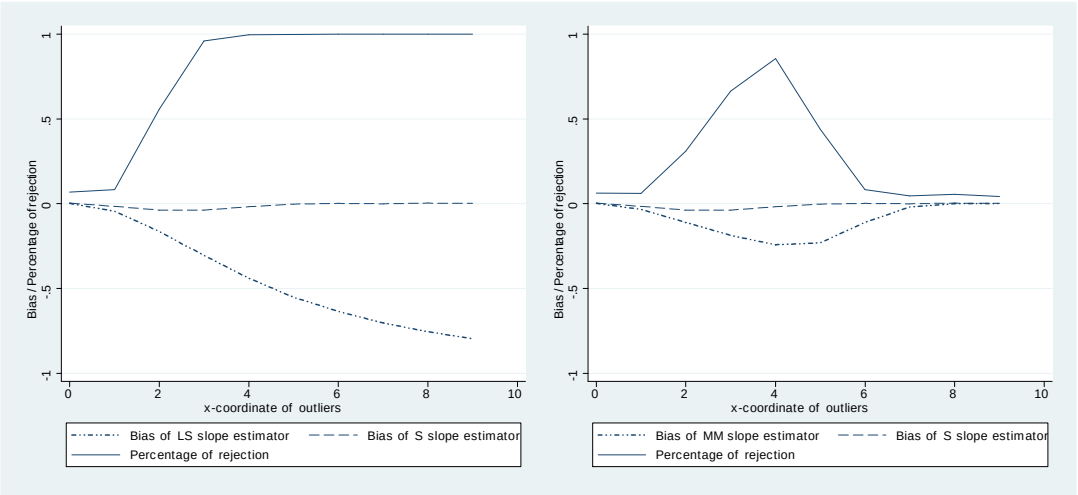

Figure 1: Bias of the estimators

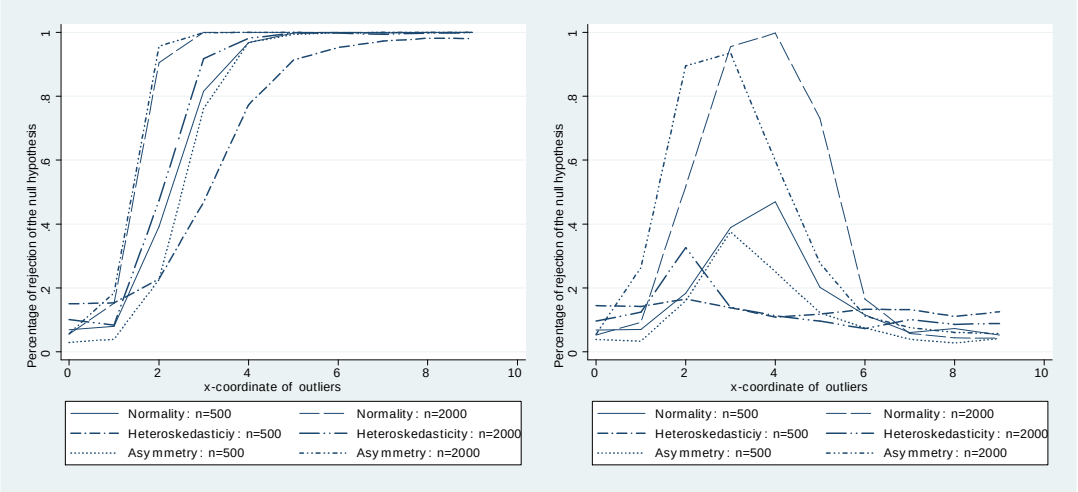

Figure 2: Power of the test.

Table 1: Percentage of rejections without contamination at $\alpha = 5\%$

		LS versus S		MM95 versus S	
Size of the test with $\alpha = 5\%$		H	YSZ	H	YSZ
i.i.d. normal errors	$n = 500$	6.08	3.12	8.86	6.60
	$n = 1000$	4.38	2.66	6.72	5.88
	$n = 2000$	5.74	5.38	5.64	5.54
Heteroskedastic errors	$n = 500$	5.30	0.12	11.80	21.30
	$n = 1000$	4.46	0.80	11.06	23.50
	$n = 2000$	8.22	8.48	7.84	21.64
Asymmetric errors	$n = 500$	4.90	99.98	4.04	100
	$n = 1000$	4.62	100	4.30	100
	$n = 2000$	4.68	100	4.40	100

Table 2: Percentage of rejection with 5% of bad leverage points

Normality		0	1	2	3	4	5	6	7	8	9
$n = 500$	LS	6.9	8.0	39.2	81.6	96.7	99.8	100	99.8	100	100
	MM	6.8	7.0	18.4	38.9	46.9	20.2	11.6	6.1	7.3	5.3
$n = 1000$	LS	6.9	8.3	55.7	96.1	99.7	99.9	100	100	100	100
	MM	6.3	6.1	30.8	66.5	85.6	44.0	8.4	4.7	5.6	4.4
$n = 2000$	LS	5.6	15.2	90.5	100	100	100	100	100	100	100
	MM	5.3	9.2	51.6	95.5	99.8	73.0	16.5	5.8	4.4	4.3
Heteroscedastic errors		0	1	2	3	4	5	6	7	8	9
$n = 500$	LS	15.0	15.3	22.8	46.8	77.3	91.3	95.2	97.2	98.2	98.1
	MM	14.4	14.2	16.5	13.8	10.9	11.8	13.3	13.2	11.1	12.5
$n = 1000$	LS	11.1	10.9	41.6	79.8	95.8	97.8	98.9	99.0	99.7	100
	MM	11.2	12.7	31.8	7.3	9.3	11.3	9.7	11.9	10.0	9.2
$n = 2000$	LS	10.1	8.4	47.3	91.7	98.2	99.8	99.8	99.4	99.8	99.9
	MM	9.6	12.4	32.6	13.9	11.3	9.6	7.3	10.1	8.6	8.9
Asymmetry errors		0	1	2	3	4	5	6	7	8	9
$n = 500$	LS	3.0	4.0	22.6	76.1	96.8	99.4	99.9	99.9	99.9	100
	MM	3.9	3.4	15.9	37.6	25.1	12.2	7.5	4.0	2.8	4.2
$n = 1000$	LS	4.8	10.5	68.0	98.7	99.8	100	100	100	100	100
	MM	4.5	15.8	57.5	61.4	32.3	15.2	7.8	5.3	4.8	5.9
$n = 2000$	LS	5.7	18.5	95.6	99.9	100	100	99.9	100	100	100
	MM	5.5	26.4	89.5	93.5	59.7	27.7	11.2	7.6	6.1	5.7

Table 3: Bias of the LS-, MM- and S-slope estimator with 5% of bad leverage points

Normality		0	1	2	3	4	5	6	7	8	9
$n = 500$	LS	-0.01	-0.06	-0.20	-0.35	-0.49	-0.60	-0.68	-0.74	-0.80	-0.83
	MM	-0.01	-0.05	-0.13	-0.20	-0.25	-0.24	-0.14	-0.03	-0.01	-0.01
	S	0.00	-0.03	-0.05	-0.04	-0.02	-0.01	0.00	0.00	-0.01	0.00
$n = 1000$	LS	0.00	-0.05	-0.17	-0.31	-0.44	-0.56	-0.64	-0.71	-0.76	-0.80
	MM	0.00	-0.04	-0.12	-0.19	-0.25	-0.23	-0.12	-0.02	0.00	0.00
	S	0.00	-0.02	-0.04	-0.04	-0.02	-0.01	0	-0.01	0.00	0.00
$n = 2000$	LS	0.00	-0.05	-0.17	-0.31	-0.45	-0.56	-0.65	-0.71	-0.77	-0.81
	MM	0.00	-0.04	-0.11	-0.18	-0.22	-0.18	-0.06	-0.01	-0.01	0.00
	S	0.00	-0.02	-0.04	-0.04	-0.01	0.00	0.00	-0.01	0.00	-0.01
Heteroscedastic errors		0	1	2	3	4	5	6	7	8	9
$n = 500$	LS	0.00	-0.03	-0.13	-0.27	-0.40	-0.51	-0.60	-0.67	-0.72	-0.77
	MM	0.00	-0.05	-0.12	-0.12	-0.09	-0.06	-0.04	-0.02	-0.01	0.00
	S	0.00	-0.03	-0.03	-0.03	-0.03	-0.02	-0.02	-0.01	-0.01	-0.01
$n = 1000$	LS	-0.01	-0.06	-0.19	-0.34	-0.48	-0.59	-0.67	-0.74	-0.79	-0.83
	MM	-0.01	-0.08	-0.16	-0.10	-0.04	-0.02	-0.01	-0.01	-0.01	-0.01
	S	0.00	-0.05	-0.02	-0.02	-0.01	-0.01	-0.01	0.00	0.00	-0.01
$n = 2000$	LS	-0.01	-0.06	-0.17	-0.32	-0.45	-0.56	-0.65	-0.72	-0.77	-0.81
	MM	-0.01	-0.07	-0.13	-0.08	-0.04	-0.03	-0.01	-0.01	-0.01	-0.01
	S	-0.01	-0.03	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Asymmetry errors		0	1	2	3	4	5	6	7	8	9
$n = 500$	LS	-0.01	-0.03	-0.12	-0.26	-0.39	-0.50	-0.59	-0.66	-0.72	-0.76
	MM	0.00	-0.01	-0.07	-0.11	-0.13	-0.12	-0.09	-0.05	-0.03	-0.02
	S	0.00	-0.01	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01
$n = 1000$	LS	-0.01	-0.05	-0.17	-0.30	-0.44	-0.55	-0.64	-0.71	-0.76	-0.80
	MM	0.00	-0.03	-0.08	-0.10	-0.08	-0.05	-0.03	-0.02	-0.02	-0.01
	S	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
$n = 2000$	LS	0.00	-0.06	-0.19	-0.34	-0.47	-0.58	-0.67	-0.74	-0.79	-0.83
	MM	-0.01	-0.03	-0.08	-0.10	-0.08	-0.05	-0.03	-0.02	-0.01	-0.01
	S	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01