Computational Statistics and Data Analysis 95 (2016) 176–191

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Case deletion diagnostics for GMM estimation

Lei Shi [a,c,*], Jun Lu [a], Jianhua Zhao [a], Gemai Chen [b]

[a] Statistics and Mathematics School, Yunnan University of Finance and Economics, Kunming, 650221, China

[b] Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, T2N 1N4 Canada

[c] Yunnan TongChuang Scientific Computing and Data Mining Center, Kunming, 650221, China

ARTICLE INFO

Article history:
Received 7 November 2014
Received in revised form 29 September 2015
Accepted 5 October 2015
Available online 22 October 2015

Keywords:
GMM estimator
Influential observations
Case deletion
Diagnostic measures
GMM-residual
GMM-leverage

ABSTRACT

Generalized method of moment (GMM) is an important estimation method for econometric models. However, it is highly sensitive to the outliers and influential observations. This paper studies the detection of influential observations using GMM estimation and establishes some useful diagnostic tools, such as residual and leverage measures. The case deletion technique is employed to derive diagnostic measures. Under linear moment conditions, an exact deletion formula is derived, and under nonlinear moment condition an approximate formula is suggested. The results are applied to efficient instrumental variable estimation and dynamic panel data models. In addition, generalized residuals and leverage measure for GMM estimator are defined and discussed. Two real data sets are used for illustration and a simulation study is conducted to confirm the usefulness of the proposed methodology.

1. Introduction

Detection of influential observations in a given data set is an important issue in statistical model building. For ordinary linear regression model, this topic has been extensively studied and results are summarized in the books of Belsley et al. (1980), Cook and Weisberg (1982, 1999), Chatterjee and Hadi (1988) and Atkinson and Riani (2000). Typically there are two approaches in detecting influential observations: case-deletion and local influence. The case deletion approach is a commonly used method and its idea is intuitive and clear, but for some complex models this method suffers from the difficulty of deriving the diagnostics. The local influence approach was first introduced by Cook (1986) and has been applied to many models because the approach can perturb different parts of the models and allow us to study local influence caused by these perturbations; see Beckman et al. (1987), Thomas and Cook (1990), Shi (1997), Poon and Poon (1999), Zhu and Lee (2001), Shi and Huang (2011) and Zhu et al. (2007).

Recently identification of influential observations has been studied for some complex models with correlated errors or mixed effects. For example, Martin (1992) suggested several influence measures for general linear models with correlated errors. Christensen et al. (1992) developed an updating formula to study the influence of observations in mixed models. Banerjee and Frees (1997) studied influence diagnostics based on subject deletion in linear longitudinal models. Haslett (1999) suggested a simpler case-deletion measure using marginal and conditional residuals. Hodges (1998) studied a case influence measure in hierarchical models by reformulating this model and approximated a case deletion formula. Haslett and Dillane (2004) developed a case deletion diagnostic for estimating the variance components in linear mixed models. Shi and Chen (2009) gave a general formula of some diagnostic measures in general linear model with correlated error. Shi and

* Corresponding author at: Statistics and Mathematics School, Yunnan University of Finance and Economics, Kunming, 650221, China.
E-mail address: shi_lei65@hotmail.com (L. Shi).

http://dx.doi.org/10.1016/j.csda.2015.10.003

Chen (2012) studied the equivalence or non-equivalence between mean-shift, replacement and deletion models in linear mixed model. Shi and Ojeda (2004) and Shi and Chen (2008) studied influence diagnostic in multilevel models.

For many econometric models, the generalized method of moment (GMM) estimator is an important estimation method (Wooldridge, 2002; Hsiao, 2003; Greene, 2003; Arellano, 2003). It has been widely used in instrumental variable models, panel data models and autoregressive conditional heteroscedasticity (ARCH) models (Engle, 1982). However, it is well known that GMM estimator is very sensitive to the outliers or influential observations (Ronchetti and Trojani, 2001). Such outliers or influential observations might potentially affect the output of parameter estimation and statistical inference, and possibly an incorrect conclusion might be obtained if such anomalous observations are not paid attention to. Therefore the detection of influential observations for GMM estimators is a crucial work. However, to the best of our knowledge there are no published measures in the literature relating to this issue.

Case deletion method detects influential observations by employing the deletion technique that derives updating formulae for some interesting parameter estimators, and the influence magnitude of a subset of observations can be monitored by comparing the parameter estimators under the full data with the data exclude this subset. In linear regression models, case deletion formulae are well developed and some important diagnostic, such as standardized residual and leverage measure also have been deeply studied and incorporated into some well-known statistical softwares. For GMM estimation, it is surprising that these issues have not been explored yet. This paper aims to establish the case deletion diagnostic method for GMM estimation. We derive the case deletion formula of GMM estimator and define residual and leverage measure in the GMM estimation framework which are different from that in traditional regression models.

The rest of the paper is organized as below: In Section 2 we derive the deletion formula for GMM estimators in an unified framework. Under linear moment condition this formula is exact and under nonlinear moment condition this formula is an approximation. In Sections 3 and 4 we apply the results of Section 2 to instrumental variable models and dynamic panel data models respectively. At the meantime, the generalized residual and leverage measure in the framework of GMM estimator are defined and discussed. In Section 5 we present two examples and a simulation study to show that the proposed methodology is effective in identifying influential observations. Proofs of the main theorems are given in the Appendix.

2. GMM estimation and deletion formulae

Consider the moment condition

$$E(g(w_i,\theta)) = E(g_i(\theta)) = 0, \tag{1}$$

where $g(\cdot,\cdot)\colon R^p \times R^k \to R^q$, w_i is a p-vector of observed variables for the ith observation (or subject) $(i = 1,\ldots,N)$, and θ is a k-vector of unknown parameter, with $k < q$. A generalized method of moment (GMM) estimator of θ, denoted by $\hat\theta$, is obtained (Hansen, 1982) by minimizing

$$\left[N^{-1}\sum_{i=1}^{N} g_i(\theta)\right]^{\top} W_N(\hat\theta_1)\left[N^{-1}\sum_{i=1}^{N} g_i(\theta)\right], \tag{2}$$

with respect to θ, where $W_N(\theta)$ satisfies that $plim_{N\to+\infty} W_N(\theta) = W_0$ with W_0 being a positive definite $q\times q$ symmetric matrix, and $\hat\theta_1$ is an initial consistent estimator of θ. Under appropriate regularity conditions (Hansen, 1982), the GMM estimator $\hat\theta$ of θ is strongly consistent and asymptotically normally distributed. The best weight matrix which yields a smaller asymptotic covariance matrix is given by

$$W_N(\hat\theta_1) = \left[N^{-1}\sum_{i=1}^{N} g_i(\hat\theta_1) g_i^{\top}(\hat\theta_1)\right]^{-1}, \tag{3}$$

and this efficient two-stage GMM estimator $\hat\theta$ has the asymptotic covariance matrix

$$V(\theta) = N^{-1}[\Gamma^{\top}(\theta) W_0 \Gamma(\theta)]^{-1}, \tag{4}$$

where $\Gamma(\theta) = E_\theta[\partial g(w_i,\theta)/\partial\theta]$, $W_0^{-1} = E_\theta[g(w_i,\theta)g^{\top}(w_i,\theta)]$. Let $C_i(\theta) = \partial g_i(\theta)/\partial\theta$, then the optimal solution to the GMM estimation can be solved by

$$\left[N^{-1}\sum_{i=1}^{N} C_i(\theta)\right]^{\top} W_N(\hat\theta_1)\left[N^{-1}\sum_{i=1}^{N} g_i(\theta)\right] = 0. \tag{5}$$

In many econometric applications, $g_i(\theta)$ may be a linear function of θ, a nonlinear function of θ, or a mixture of the two (Ahn and Schmidt, 1997). However, it is known that $g_i(\theta) = g(\omega_i,\theta)$ in many cases are unbounded with respect to the observations (Ronchetti and Trojani, 2001) thus the GMM estimator is highly sensitive for the presence of outliers or influential observations. Therefore it is extremely important to study the detection of observations that have potential influence on the GMM estimators. This research is sometimes referred to as sensitivity analysis (Chatterjee and Hadi, 1988).