A Finite Sample Correction for the Variance of Linear Two-Step GMM Estimators

Frank Windmeijer

THE INSTITUTE FOR FISCAL STUDIES
WP 00/19

A Finite Sample Correction for the Variance of Linear Two-Step GMM Estimators*

Frank Windmeijer
Institute for Fiscal Studies
7 Ridgmount Street
London WC1E 7AE
f.windmeijer@ifs.org.uk

Abstract

Monte Carlo studies have shown that estimated asymptotic standard errors of the efficient two-step generalised method of moments (GMM) estimator can be severely downward biased in small samples. The weight matrix used in the calculation of the efficient two-step GMM estimator is based on initial consistent parameter estimates. In this paper it is shown that the extra variation due to the presence of these estimated parameters in the weight matrix accounts for much of the difference between the finite sample and the asymptotic variance of the two-step GMM estimator that utilises moment conditions that are linear in the parameters. This difference can be estimated, resulting in a finite sample corrected estimate of the variance. In a Monte Carlo study of a panel data model it is shown that the corrected variance estimate approximates the finite sample variance well, leading to more accurate inference.

Key Words: Generalised Method of Moments; Variance correction; Panel data

JEL Classification: C12, C23

---

*I am grateful to Manuel Arellano, Richard Blundell and Stephen Bond for helpful comments. Financial support of the ESRC, grant no. R000223068, is gratefully acknowledged. This work has benefited from a longer term programme of research supported by the ESRC Centre for the Microeconomic Analysis of Fiscal Policy at IFS. The usual disclaimer applies.

1. Introduction

In Monte Carlo studies it has often been found that the estimated asymptotic standard errors of the efficient two-step generalized method of moments (GMM) estimator are severely downward biased in small samples,[1] see e.g. Arellano and Bond (1991), whereas the asymptotic standard errors of one-step GMM estimators are virtually unbiased. One-step GMM estimators use weight matrices that are independent of estimated parameters, whereas the efficient two-step GMM estimator weighs the moment conditions by a consistent estimate of their covariance matrix. This weight matrix is constructed using an initial consistent estimate of the parameters in the model. In this paper it is shown that the extra variation due to the presence of these estimated parameters in the weight matrix accounts for much of the difference between the finite sample and the asymptotic variance of the two-step GMM estimator that utilizes moment conditions that are linear in the parameters. This difference can be estimated, resulting in finite sample corrected estimates of the variance. In a Monte Carlo study of a panel data model, it is shown that this corrected variance approximates the finite sample variance of the two-step GMM estimator well, leading to more accurate inference. The variance correction is further illustrated using the models and data from Arellano and Bond (1991) and Blundell and Bond (2000).

In section 2 the finite sample bias of the asymptotic variance of the two-step

[1] The same observation has been made for alternative GMM estimators, like the continuously updated and iterated GMM estimators (see Hansen, Heaton and Yaron (1996)). A finite sample variance correction for the iterated GMM estimator is discussed in section 2.1.

GMM estimator is derived. Section 3 considers a bivariate panel data model, and Section 4 presents Monte Carlo results for this model. Section 5 presents the empirical applications and section 6 concludes. Further, the finite sample performance of the Sargan/Hansen test for overidentifying restrictions is discussed in the appendix.

2. GMM and Finite Sample Variance Correction

Consider the moment conditions

$$E\left[g\left(X_i, \theta_0\right)\right] = E\left[g_i\left(\theta_0\right)\right] = 0,$$

where $g\left(.\right)$ is vector of order q and θ_0 is a parameter vector of order k, with $k < q$. The GMM estimator $\widehat{\theta}$ for θ_0 minimizes[2]

$$Q_{W_N} = \left[\frac{1}{N}\sum_{i=1}^{N} g_i\left(\theta\right)\right]' W_N^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} g_i\left(\theta\right)\right],$$

with respect to θ; where W_N is a positive semidefinite matrix which satisfies $\text{plim}_{N\to\infty} W_N = W$, with W a positive definite matrix. Regularity conditions are assumed such that $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} g_i\left(\theta\right) = E\left[g_i\left(\theta\right)\right]$ and $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i\left(\theta_0\right) \to N\left(0, \Psi\right)$. Let $\Gamma\left(\theta\right) = E\left[\partial g_i\left(\theta\right)/\partial\theta'\right]$ and $\Gamma_{\theta_0} \equiv \Gamma\left(\theta_0\right)$, then $\sqrt{N}\left(\widehat{\theta} - \theta_0\right)$ has a limiting normal distribution, $\sqrt{N}\left(\widehat{\theta} - \theta_0\right) \to N\left(0, V_W\right)$, where

$$V_W = \left(\Gamma'_{\theta_0} W^{-1} \Gamma_{\theta_0}\right)^{-1} \Gamma'_{\theta_0} W^{-1} \Psi W^{-1} \Gamma_{\theta_0} \left(\Gamma'_{\theta_0} W^{-1} \Gamma_{\theta_0}\right)^{-1}. \qquad (2.1)$$

The efficient two-step GMM estimator, denoted $\widehat{\theta}_2$, is based on a weight matrix that satisfies $\text{plim}_{N\to\infty} W_N = \Psi$, with $V_W = \left(\Gamma'_{\theta_0} \Psi^{-1} \Gamma_{\theta_0}\right)^{-1}$. A weight matrix that

[2] See Hansen (1982).

satisfies this property is given by

$$W_N\left(\widehat{\theta}_1\right)=\frac{1}{N}\sum_{i=1}^{N}g_i\left(\widehat{\theta}_1\right)g_i\left(\widehat{\theta}_1\right)', \tag{2.2}$$

where $\widehat{\theta}_1$ is an initial consistent estimator for θ_0.

Let

$$\begin{aligned}
\overline{g}(\theta) &= \frac{1}{N}\sum_{i=1}^{N}g_i(\theta)\\
C(\theta) &= \frac{\partial\overline{g}(\theta)}{\partial\theta'}\\
G(\theta) &= \frac{\partial C(\theta)}{\partial\theta}=\begin{bmatrix}\frac{\partial C(\theta)}{\partial\theta_1}\\ \frac{\partial C(\theta)}{\partial\theta_2}\\ \vdots\\ \frac{\partial C(\theta)}{\partial\theta_k}\end{bmatrix},
\end{aligned}$$

and

$$\begin{aligned}
b_{\theta_0,W_N} &= \frac{\partial Q_{W_N}}{\partial\theta}|_{\theta_0}=C(\theta_0)'W_N^{-1}\overline{g}(\theta_0);\\
A_{\theta_0,W_N} &= \frac{\partial^2 Q_{W_N}}{\partial\theta\partial\theta'}|_{\theta_0}=C(\theta_0)'W_N^{-1}C(\theta_0)+G(\theta_0)'\left(I_k\otimes W_N^{-1}\overline{g}(\theta_0)\right)
\end{aligned}$$

A standard first order Taylor series approximation of $\widehat{\theta}_2$ around θ_0, conditional on $W_N\left(\widehat{\theta}_1\right)$, results in

$$\widehat{\theta}_2-\theta_0=-A^{-1}_{\theta_0,W_N\left(\widehat{\theta}_1\right)}b_{\theta_0,W_N\left(\widehat{\theta}_1\right)}+O_p\left(N^{-1}\right),$$

and an estimate for the asymptotic variance of $\widehat{\theta}_2$ is given by

$$\widehat{\text{var}}\left(\widehat{\theta}_2\right)=\frac{1}{N}A^{-1}_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}C\left(\widehat{\theta}_2\right)'W_N^{-1}\left(\widehat{\theta}_1\right)C\left(\widehat{\theta}_2\right)A^{-1}_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}.$$

However, a further expansion of $\widehat{\theta}_1$ around θ_0 results in

$$\widehat{\theta}_2-\theta_0=-A^{-1}_{\theta_0,W_N(\theta_0)}b_{\theta_0,W_N(\theta_0)}+D_{\theta_0,W_N(\theta_0)}\left(\widehat{\theta}_1-\theta_0\right)+O_p\left(N^{-1}\right),$$

where

$$W_N(\theta_0) = \frac{1}{N}\sum_{i=1}^{N} g_i(\theta_0)\, g_i(\theta_0)'$$

and

$$D_{\theta_0, W_N(\theta_0)} = \frac{\partial}{\partial \theta'}\left(-A^{-1}_{\theta_0, W_N(\theta)} b_{\theta_0, W_N(\theta)}\right)|_{\theta_0}$$

is a $k \times k$ matrix. The j-th column of $D_{\theta_0, W_N(\theta_0)}$ is given by[3]

$$\begin{aligned} & - A^{-1}_{\theta_0, W_N(\theta_0)} C(\theta_0)' W_N^{-1}(\theta_0) \frac{\partial W_N(\theta)}{\partial \theta_j}|_{\theta_0} W_N^{-1}(\theta_0) C(\theta_0) A^{-1}_{\theta_0, W_N(\theta_0)} b_{\theta_0, W_N(\theta_0)} \qquad (2.3) \\ & - A^{-1}_{\theta_0, W_N(\theta_0)} G(\theta_0)' \left(I_k \otimes W_N^{-1}(\theta_0) \frac{\partial W_N(\theta)}{\partial \theta_j}|_{\theta_0} W_N^{-1}(\theta_0) \overline{g}(\theta_0)\right) A^{-1}_{\theta_0, W_N(\theta_0)} b_{\theta_0, W_N(\theta_0)} \\ & + A^{-1}_{\theta_0, W_N(\theta_0)} C(\theta_0)' W_N^{-1}(\theta_0) \frac{\partial W_N(\theta)}{\partial \theta_j}|_{\theta_0} W_N^{-1}(\theta_0) \overline{g}(\theta_0), \end{aligned}$$

where

$$\frac{\partial W_N(\theta)}{\partial \theta_j} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial g_i(\theta)}{\partial \theta_j} g_i(\theta)' + g_i(\theta) \frac{\partial g_i(\theta)'}{\partial \theta_j}\right).$$

The first two terms of $D_{\theta_0, W_N(\theta_0)}$ are functions of $A^{-1}_{\theta_0, W_N(\theta_0)} b_{\theta_0, W_N(\theta_0)}$ which is the bias of an infeasible GMM estimator that uses an efficient weight matrix that is based on the true parameters θ_0. This bias tends to be small and will generally not grow with the number of instruments, see Newey and Smith (2000). The third term, which in general does increase with the number of moment conditions, will therefore dominate.

Taking $\widehat{\theta}_1$ as a one-step GMM estimator using a weight matrix that does not depend on estimated parameters:

$$\widehat{\theta}_1 - \theta_0 = -A^{-1}_{\theta_0, W_N} b_{\theta_0, W_N} + O_p\left(N^{-1}\right),$$

[3] Using results as given in Magnus and Neudecker (1988, p.151).

in practice, an estimate of the variance of $\widehat{\theta}_2$ that incorporates the term involving the one-step estimator can be obtained as

$$\begin{aligned}
\widehat{\text{var}}_c\left(\widehat{\theta}_2\right) &= \frac{1}{N}A^{-1}_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}C\left(\widehat{\theta}_2\right)'W_N^{-1}\left(\widehat{\theta}_1\right)C\left(\widehat{\theta}_2\right)A^{-1}_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)} \\
&\quad + D_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}\widehat{\text{var}}\left(\widehat{\theta}_1\right)D'_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)} \\
&\quad + \frac{1}{N}D_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}A^{-1}_{\widehat{\theta}_1,W_N}C\left(\widehat{\theta}_1\right)'W_N^{-1}C\left(\widehat{\theta}_2\right)A^{-1}_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)} \\
&\quad + \frac{1}{N}A^{-1}_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}C\left(\widehat{\theta}_2\right)'W_N^{-1}C\left(\widehat{\theta}_1\right)A^{-1}_{\widehat{\theta}_1,W_N}D'_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}.
\end{aligned} \tag{2.4}$$

where the estimated variance of the one-step estimator is given by

$$\widehat{\text{var}}\left(\widehat{\theta}_1\right) = \frac{1}{N}A^{-1}_{\widehat{\theta}_1,W_N}C\left(\widehat{\theta}_1\right)'W_N^{-1}W_N\left(\widehat{\theta}_1\right)W_N^{-1}C\left(\widehat{\theta}_1\right)A^{-1}_{\widehat{\theta}_1,W_N}$$

and $D_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}$ is as defined in (2.3) with θ_0 and $W_N\left(\theta_0\right)$ substituted by $\widehat{\theta}_2$ and $W_N\left(\widehat{\theta}_1\right)$ respectively. The first two terms of $D_{\widehat{\theta}_2,W_N\left(\widehat{\theta}_1\right)}$ are equal to zero.

The term $D_{\theta_0,W(\theta_0)}\left(\widehat{\theta}_1-\theta_0\right)$ is itself $O_p\left(N^{-1}\right)$ and in this general setting, incorporating non-linear models and/or non-linear moment conditions, whether taking account of it will improve small sample inference depends on the other remainder terms which are of the same order.

A definite improvement, however, will be obtained in models where all the moment conditions used are linear in the parameters, as in this case

$$\begin{aligned}
\widehat{\theta}_{2l}-\theta_{0l} &= -\left(C'W_N^{-1}\left(\widehat{\theta}_{1l}\right)C\right)^{-1}C'W_N^{-1}\left(\widehat{\theta}_{1l}\right)\overline{g}\left(\theta_{0l}\right) \\
&= -\left(C'W_N^{-1}\left(\theta_{0l}\right)C\right)^{-1}C'W_N^{-1}\left(\theta_{0l}\right)\overline{g}\left(\theta_{0l}\right) \\
&\quad + D_{\theta_{0l},W_N(\theta_{0l})}\left(\widehat{\theta}_{1l}-\theta_{0l}\right) + O_p\left(N^{-\frac{3}{2}}\right),
\end{aligned}$$

where the subscript $_l$ indicates parameters in linear moment conditions and the j-th column of $D_{\theta_{0l},W_N(\theta_{0l})}$ is given by

$$
\begin{aligned}
& -\left(C'W_N^{-1}(\theta_{0l})\,C\right)^{-1} C'W_N^{-1}(\theta_{0l})\,\frac{\partial W_N(\theta_l)}{\partial \theta_{lj}}|_{\theta_{0l}} W_N^{-1}(\theta_{0l})\,C \\
& \qquad\qquad \times \left(C'W_N^{-1}(\theta_{0l})\,C\right)^{-1} C'W_N^{-1}(\theta_{0l})\,\overline{g}(\theta_{0l}) \\
& +\left(C'W_N^{-1}(\theta_{0l})\,C\right)^{-1} C'W_N^{-1}(\theta_{0l})\,\frac{\partial W_N(\theta_l)}{\partial \theta_{lj}}|_{\theta_{0l}} W_N^{-1}(\theta_{0l})\,\overline{g}(\theta_{0l})\,.
\end{aligned}
$$

Therefore, in this case the term $D_{\theta_{0l},W_N(\theta_{0l})}\left(\widehat{\theta}_{1l}-\theta_{0l}\right)$ will improve the accuracy of the approximation in finite samples.

A one-step linear estimator satisfies

$$
\widehat{\theta}_{1l}-\theta_{0l} = -\left(C'W_N^{-1}C\right)^{-1} C'W_N^{-1}\overline{g}(\theta_{0l})\,,
$$

and the finite sample corrected estimate of the variance of $\widehat{\theta}_{2l}$ can be obtained as

$$
\begin{aligned}
\widehat{\text{var}}_c\left(\widehat{\theta}_{2l}\right) \;=\;& \frac{1}{N}\left(C'W_N^{-1}\left(\widehat{\theta}_{1l}\right)C\right)^{-1} + D_{\widehat{\theta}_{2l},W_N\left(\widehat{\theta}_{1l}\right)}\widehat{\text{var}}\left(\widehat{\theta}_{1l}\right)D'_{\widehat{\theta}_{2l},W_N\left(\widehat{\theta}_{1l}\right)} \\
& + \frac{1}{N}D_{\widehat{\theta}_{2l},W_N\left(\widehat{\theta}_{1l}\right)}\left(C'W_N^{-1}\left(\widehat{\theta}_{1l}\right)C\right)^{-1} + \frac{1}{N}\left(C'W_N^{-1}\left(\widehat{\theta}_{1l}\right)C\right)^{-1}D'_{\widehat{\theta}_{2l},W_N\left(\widehat{\theta}_{1l}\right)}
\end{aligned}
$$

where the first RHS term is the conventional estimate of the asymptotic variance, and

$$
\widehat{\text{var}}\left(\widehat{\theta}_{1l}\right) = \frac{1}{N}\left(C'W_N^{-1}C\right)^{-1} C'W_N^{-1}W_N\left(\widehat{\theta}_{1l}\right)W_N^{-1}C\left(C'W_N^{-1}C\right)^{-1}.
$$

2.1. The Iterated GMM Estimator

The iterated GMM estimator, denoted $\widehat{\theta}_{IT}$, is a multi-step GMM estimator that is iterated until convergence. Therefore, in the case of linear moment conditions,

$$
\widehat{\theta}_{ITl}-\theta_{0l} = -\left(C'W_N^{-1}\left(\widehat{\theta}_{ITl}\right)C\right)^{-1} C'W_N^{-1}\left(\widehat{\theta}_{ITl}\right)\overline{g}(\theta_{0l})\,,
$$

and the asymptotic variance is estimated as

$$\widehat{\text{var}}\left(\widehat{\theta}_{ITl}\right) = \frac{1}{N}\left(C'W_N^{-1}\left(\widehat{\theta}_{ITl}\right)C\right)^{-1}.$$

Using the same arguments as above, it follows that

$$\begin{aligned} \widehat{\theta}_{ITl} - \theta_{0l} &= -\left(C'W_N^{-1}\left(\theta_{0l}\right)C\right)^{-1} C'W_N^{-1}\left(\theta_{0l}\right)\overline{g}\left(\theta_{0l}\right) \\ &\quad + D_{\theta_{0l},W_N(\theta_{0l})}\left(\widehat{\theta}_{ITl} - \theta_{0l}\right) + \ O_p\left(N^{-\frac{3}{2}}\right) \end{aligned}$$

and a corrected finite sample estimate for the variance of the iterated linear GMM estimator is given by

$$\widehat{\text{var}}_c\left(\widehat{\theta}_{ITl}\right) = \frac{1}{N}\left(I - D_{\widehat{\theta}_{ITl},W_N\left(\widehat{\theta}_{ITl}\right)}\right)^{-1}\left(C'W_N^{-1}\left(\widehat{\theta}_{ITl}\right)C\right)^{-1}\left(I - D'_{\widehat{\theta}_{ITl},W_N\left(\widehat{\theta}_{ITl}\right)}\right)^{-1}.$$

3. A Panel Data Model

Consider the panel data model specification

$$\begin{aligned} y_{it} &= \beta_0 x_{it} + u_{it} \\ u_{it} &= \eta_i + v_{it} \end{aligned}$$

for $i = 1, ..., N,\ t = 2, ..., T$. The single regressor x_{it} is correlated with η_i and predetermined with respect to v_{it}, meaning that $E\left(x_{it}v_{it+s}\right) = 0,\ s = 0, .., T - t$, but $E\left(x_{it}v_{it-r}\right) \neq 0,\ r = 1, ..., t - 1$. A commonly used estimator is the GMM estimator in the model in first differences, see Arellano and Bond (1991),

$$\Delta y_{it} = \beta_0 \Delta x_{it} + \Delta u_{it};\ \ t = 2, ..., T$$

with $T(T-1)/2$ sequential instruments

$$Z_i = \begin{bmatrix} x_{i1} & 0 & 0 & 0 & 0 & 0 \\ 0 & x_{i1} & x_{i2} & 0 & 0 & 0 \\ & & \ddots & & & \\ 0 & 0 & 0 & x_{i1} & \ldots & x_{iT-1} \end{bmatrix}.$$

The moment conditions are then given by $E(Z_i' \Delta u_i) = 0$, where Δu_i is the $(T-1)$ vector $(\Delta u_{i2}, ..., \Delta u_{iT})'$. The covariance matrix of the moment conditions is denoted Ψ.

A one-step GMM estimator is given by

$$\widehat{\beta}_1 = \left(\Delta x' Z W_N^{-1} Z' \Delta x\right)^{-1} \Delta x' Z W_N^{-1} Z' \Delta y,$$

where Z' is the $T(T-1)/2 \times N(T-1)$ matrix $(Z_1', Z_2', ..., Z_N')$, Δx_i is the $(T-1)$ vector $(\Delta x_{i2}, ..., \Delta x_{iT})'$, Δy_i is the $(T-1)$ vector $(\Delta y_{i2}, ..., \Delta y_{iT})'$, Δx and Δy are $N(T-1)$ vectors $(\Delta x_1', \Delta x_2', ..., \Delta x_N')'$ and $(\Delta y_1', \Delta y_2', ..., \Delta y_N')'$ respectively, and W_N^{-1} is an initial positive definite weight matrix. For example, 2SLS sets $W_N = \frac{1}{N} Z'Z$. An initial weight matrix that is efficient when the u_{it} are i.i.d. is $W_N = \frac{1}{N} \sum_{i=1}^N Z_i' A Z_i$, where A is a matrix with 2's on the main diagonal, -1's on the first off-diagonals and zeros elsewhere.

The asymptotic variance of $\widehat{\beta}_1$ is estimated by

$$\widehat{\text{var}}\left(\widehat{\beta}_1\right) = N\left(\Delta x' Z W_N^{-1} Z' \Delta x\right)^{-1} \Delta x' Z W_N^{-1} W_N\left(\widehat{\beta}_1\right) W_N^{-1} Z' \Delta x \left(\Delta x' Z W_N^{-1} Z' \Delta x\right)^{-1},$$

where

$$\begin{aligned} W_N\left(\widehat{\beta}_1\right) &= \frac{1}{N} \sum_{i=1}^N Z_i' \Delta \widehat{u}_{i1} \Delta \widehat{u}_{i1}' Z_i \\ \Delta \widehat{u}_{i1} &= \Delta y_i - \widehat{\beta}_1 \Delta x_i \end{aligned}$$

with $W_N\left(\widehat{\beta}_1\right)$ a consistent estimate of Ψ. Given the estimate $\widehat{\beta}_1$, the efficient two-step GMM estimator is given by

$$\widehat{\beta}_2=\left(\Delta x'ZW_N^{-1}\left(\widehat{\beta}_1\right)Z'\Delta x\right)^{-1}\Delta x'ZW_N^{-1}\left(\widehat{\beta}_1\right)Z'\Delta y$$

Standard theory implies that the asymptotic variance of $\widehat{\beta}_2$ is estimated by

$$\widehat{\text{var}}\left(\widehat{\beta}_2\right)=N\left(\Delta x'ZW_N^{-1}\left(\widehat{\beta}_1\right)Z'\Delta x\right)^{-1}, \tag{3.1}$$

which is an estimate of $\frac{1}{N}\left(\Gamma'_{Z\Delta x}\Psi^{-1}\Gamma_{Z\Delta x}\right)^{-1}$, with $\Gamma_{Z\Delta x}=\text{plim}_{N\to\infty}\frac{1}{N}Z'\Delta x$.

It is well documented (see for example Arellano and Bond (1991)) that the estimated standard errors of $\widehat{\beta}_2$ are downward biased in small samples, leading to a very poor performance of the Wald test. Applying the Taylor series expansion developed in the previous section to account for the presence of $\widehat{\beta}_1$ in the estimated weight matrix results in

$$\begin{aligned}\widehat{\beta}_2-\beta_0 &= \left(\Delta x'ZW_N^{-1}\left(\beta_0\right)Z'\Delta x\right)^{-1}\Delta x'ZW_N^{-1}\left(\beta_0\right)Z'\Delta u \\ &\quad + D_{\beta_0,W_N(\beta_0)}\left(\widehat{\beta}_1-\beta_0\right)+O_p\left(N^{-\frac{3}{2}}\right),\end{aligned} \tag{3.2}$$

where Δu is the $N\left(T-1\right)$ vector $\left(\Delta u_1',\Delta u_2',...,\Delta u_N'\right)'$, and $D_{\beta_0,W_N(\beta_0)}$ is given by

$$\begin{aligned}D_{\beta_0,W_N(\beta_0)} &= \left(\Delta x'ZW_N^{-1}\left(\beta_0\right)Z'\Delta x\right)^{-1}\Delta x'ZW_N^{-1}\left(\beta_0\right)\frac{\partial W_N\left(\beta\right)}{\partial\beta}|_{\beta_0}W_N^{-1}\left(\beta_0\right)Z'\Delta x \\ &\quad \times\left(\Delta x'ZW_N^{-1}\left(\beta_0\right)Z'\Delta x\right)^{-1}\Delta x'ZW_N^{-1}\left(\beta_0\right)Z'\Delta u \\ &\quad -\left(\Delta x'ZW_N^{-1}\left(\beta_0\right)Z'\Delta x\right)^{-1}\Delta x'ZW_N^{-1}\left(\beta_0\right)\frac{\partial W_N\left(\beta\right)}{\partial\beta}|_{\beta_0}W_N^{-1}\left(\beta_0\right)Z'\Delta u,\end{aligned}$$

with

$$W_N\left(\beta_0\right)=\frac{1}{N}\sum_{i=1}^{N}Z_i'\Delta u_i\Delta u_i'Z_i$$

and

$$\frac{\partial W_N(\beta)}{\partial \beta}|_{\beta_0} = -\frac{1}{N}\sum_{i=1}^{N} Z_i' \left(\Delta x_i \Delta u_i' + \Delta u_i \Delta x_i'\right) Z_i.$$

A small sample bias corrected estimate of the variance of $\widehat{\beta}_2$ can then be obtained as[4]

$$\begin{aligned}\widehat{\text{var}}_c\left(\widehat{\beta}_2\right) &= N\left(\Delta x' Z W_N^{-1}\left(\widehat{\beta}_1\right) Z'\Delta x\right)^{-1} + D_{\widehat{\beta}_2, W_N\left(\widehat{\beta}_1\right)} \widehat{\text{var}}\left(\widehat{\beta}_1\right) D'_{\widehat{\beta}_2, W_N\left(\widehat{\beta}_1\right)} \\ &\quad + N D_{\widehat{\beta}_2, W_N\left(\widehat{\beta}_1\right)}\left(\Delta x' Z W_N^{-1}\left(\widehat{\beta}_1\right) Z'\Delta x\right)^{-1} \\ &\quad + N\left(\Delta x' Z W_N^{-1}\left(\widehat{\beta}_1\right) Z'\Delta x\right)^{-1} D'_{\widehat{\beta}_2, W_N\left(\widehat{\beta}_1\right)}\end{aligned} \tag{3.3}$$

Again, as $\left(\Delta x' Z W_N^{-1}\left(\widehat{\beta}_1\right) Z'\Delta x\right)^{-1} \Delta x' Z W_N^{-1}\left(\widehat{\beta}_1\right) Z'\Delta \widehat{u}_2 = 0$, where $\Delta \widehat{u}_2 = \Delta y_i - \widehat{\beta}_2 \Delta x$, the expression of $D_{\widehat{\beta}_2, W_N\left(\widehat{\beta}_1\right)}$ simplifies.

4. Monte Carlo Results

A panel data process is generated as:

$$\begin{aligned} y_{it} &= \beta_0 x_{it} + \eta_i + v_{it} \\ x_{it} &= 0.5 x_{it-1} + \eta_i + 0.5 v_{it-1} + \varepsilon_{it} \\ \eta_i &\sim N(0,1) \qquad \varepsilon_{it} \sim N(0,1) \\ v_{it} &= \delta_i \tau_t \omega_{it} \qquad \omega_{it} \sim \left(\chi_1^2 - 1\right) \\ \delta_i &\sim U[0.5, 1.5] \qquad \tau_t = 0.5 + 0.1(t-1) \end{aligned}$$

Fifty time periods are generated, with $\tau_t = 0.5$ for $t = -49, ..., 0$ and $x_{i,-49} \sim N\left(\frac{\eta_i}{0.5} + \frac{1}{0.75}\right)$, before the estimation sample is drawn. This model design corresponds to the features of the panel data model described in the previous section,

[4] Note that $D_{\widehat{\beta}_2, W_N\left(\widehat{\beta}_1\right)}$ is a scalar here.

the x_{it} are correlated with the unobserved heterogeneity η_i and are predetermined. The design is further such that the v_{it} are skewed and heteroscedastic over both time t and individual i. The parameters are estimated by first differenced GMM as described in the previous section.

Table 1 reports estimation results for $\beta_0 = 1$, $N = 100$, $T = 4$ and $T = 8$. Reported are means and standard deviations of a one-step GMM estimator with $W_N = \frac{1}{N}\sum_{i=1}^{N} Z_i' A Z_i$, which is not efficient in this case, the two-step GMM estimator using $W_N\left(\widehat{\beta}_1\right)$, and an infeasible estimator that uses the true parameter β_0 to evaluate the weight matrix, $W_N(\beta_0)$. This latter estimator is denoted $\widehat{\beta}_{W_N(\beta_0)}$. For all three GMM estimators, means of the conventional asymptotic standard errors are reported, denoted $se\ \widehat{\beta}_1$, $se\ \widehat{\beta}_2$, and $se\ \widehat{\beta}_{W_N(\beta_0)}$. For the two-step GMM estimator the components of the Taylor series expansion (3.2) are calculated directly at the true value of β_0, and the standard deviation of the sum of the two infeasible components is denoted $sd_{\inf}\ \widehat{\beta}_2$. The means of the feasible estimated corrected standard errors, calculated from (3.3), is denoted $sec\ \widehat{\beta}_2$.

The estimated asymptotic standard errors of the GMM estimators that do not have estimated parameters present in the weight matrix, $\widehat{\beta}_1$ and $\widehat{\beta}_{W_N(\beta_0)}$, are on average only slightly smaller than their standard deviations, less so at $T = 8$ than at $T = 4$. For the two-step GMM estimator, however, the means of the estimated asymptotic standard errors are considerably smaller than the standard deviations of $\widehat{\beta}_2$, especially at $T = 8$. At $T = 4$, $se\ \widehat{\beta}_2$ accounts for 87% of $sd\ \widehat{\beta}_2$, whereas when $T = 8$, $se\ \widehat{\beta}_2$ accounts for only 66% of $sd\ \widehat{\beta}_2$. When $T = 8$, there are 28 instruments, whereas there are only 6 instruments when $T = 4$.

The standard deviations of the Taylor series expansion (3.2) are almost equal to the standard deviations of $\widehat{\beta}_2$. The standard deviations of the leading term in (3.2), $\left(\Delta x' Z W_N^{-1}(\beta_0) Z' \Delta x\right)^{-1} \Delta x' Z W_N^{-1}(\beta_0) Z' \Delta u$, are given by $sd\ \widehat{\beta}_{W_N(\beta_0)}$, and so the term involving $\left(\widehat{\beta}_1 - \beta_0\right)$ accounts for 10% and 33% of the standard deviation of $\widehat{\beta}_2$ for $T = 4$ and $T = 8$ respectively. As can be seen, the conventional estimated asymptotic standard error of $\widehat{\beta}_2$, $se\ \widehat{\beta}_2$, is in fact a good estimate of the standard error of the estimator $\widehat{\beta}_{W_N(\beta_0)}$ rather than the standard error of $\widehat{\beta}_2$, and the difference between $sd\ \widehat{\beta}_2$ and $se\ \widehat{\beta}_2$ is due to the presence of the estimated $\widehat{\beta}_1$ in the weight matrix.

Table 1. Monte Carlo Results

	$T = 4$	$T = 8$
$\widehat{\beta}_1$	0.9800	0.9738
$sd\ \widehat{\beta}_1$	0.1534	0.0832
$se\ \widehat{\beta}_1$	0.1471	0.0809
$\widehat{\beta}_2$	0.9868	0.9810
$sd\ \widehat{\beta}_2$	0.1423	0.0721
$se\ \widehat{\beta}_2$	0.1244	0.0477
$sd_{\text{inf}}\ \widehat{\beta}_2$	0.1414	0.0717
$sec\ \widehat{\beta}_2$	0.1391	0.0715
$\widehat{\beta}_{W_N(\beta_0)}$	0.9895	0.9915
$sd\ \widehat{\beta}_{W_N(\beta_0)}$	0.1278	0.0481
$se\ \widehat{\beta}_{W_N(\beta_0)}$	0.1229	0.0474

Notes: $N = 100$, $\beta_0 = 1$, means and standard deviations of $10,000$ replications.
$sd_{\text{inf}}\ \widehat{\beta}_2$ is the standard deviation of the first order Taylor series expansion (3.2) evaluated at β_0.
$sec\ \widehat{\beta}_2$ is the estimated standard error of $\widehat{\beta}_2$ corrected for small sample bias.
$\widehat{\beta}_{W(\beta_0)}$ is the GMM estimator for β_0 using $W_N(\beta_0)$

The means of the feasible estimated standard errors that correct for this extra variation due to the estimation of the efficient weight matrix, as estimated from (3.3), are close to the standard deviations of $\widehat{\beta}_2$. The corrected standard errors now account for 98% and 99% of the standard deviation of $\widehat{\beta}_2$, for $T = 4$ and $T = 8$ respectively.

In order to evaluate the behavior of the Wald test statistics for the test $H_0 : \beta_0 = 1$, based on the one-step and two-step estimators and associated standard errors, Figures 1 and 2 show p-value plots (see Davidson and MacKinnon (1996)) for three Wald statistics, for $T = 4$ and $T = 8$ respectively, based on the same $10,000$ Monte Carlo replications. WALD_1 is based on the one-step estimator and its asymptotic standard error. WALD_2 is based on the conventional two-step estimation results, whereas WALD_{2C} uses the corrected variance estimate.

For $T = 4$, WALD_2 is moderately oversized, whereas WALD_1 and WALD_{2C} have good size properties. For $T = 8$, WALD_2 is severely oversized. Using the corrected standard errors improves the size of the test dramatically and WALD_{2C} is only slightly oversized. WALD_1 is more oversized than WALD_{2C} as it has a larger small sample bias. It is clear that using the corrected variance estimate for the two-step estimator improves the finite sample inference considerably.[5]

[5] The size performance of the Sargan/Hansen test for overidentifying restrictions is not affected by the presence of estimated parameters in the weight matrix. This is illustrated in the Appendix.

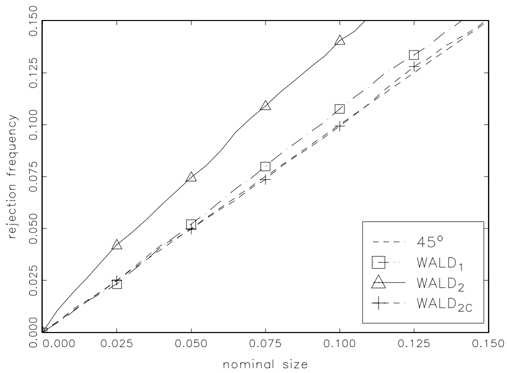

Fig 1. P-value plot, $H_0 : \beta_0 = 1$, $T = 4$

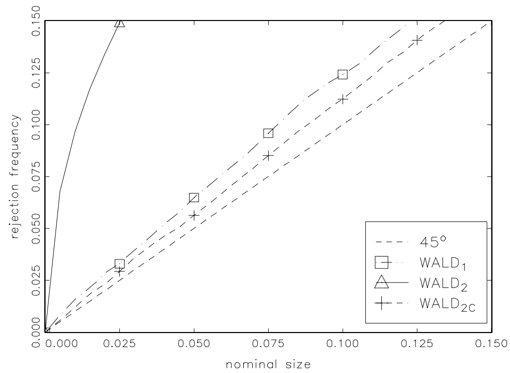

Fig 2. P-value plot, $H_0 : \beta_0 = 1$, $T = 8$

5. Empirical Illustrations

In this section results of the two-step GMM variance correction are illustrated for two examples from the literature. The first example is taken from Arellano and Bond (1991), who used a sample of 140 UK quoted firms over the years 1976-1984. The sample is unbalanced with observations varying between 7 and 9 records per company. Arellano and Bond (1991) estimated dynamic employment equations, one of which was specified as

$$n_{it} = \alpha_1 n_{it-1} + \alpha_2 n_{it-2} + \beta w_{it} + \beta_1 w_{it-1} + \gamma k_{it} + \delta ys_{it} + \delta_1 ys_{it-1} + \lambda_t + \eta_i + u_{it},$$

where n_{it} is the logarithm of UK employment in company i at the end of period t, w_{it} is the log of the real product wage, k_{it} is the log of gross capital and ys_{it} is the log of industry output. The model is estimated in first differences, with the instrument set of the form

$$Z_i = \begin{bmatrix} n_{i1} & n_{i2} & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \Delta x'_{i4} \\ 0 & 0 & n_{i1} & n_{i2} & n_{i3} & & 0 & & 0 & \Delta x'_{i5} \\ \vdots & & & & & \ddots & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & n_{i1} & \cdots & n_{i7} & \Delta x'_{i9} \end{bmatrix}$$

where $\Delta x'_{it} = [1, \Delta w_{it}, \Delta w_{it-1}, \Delta k_{it}, \Delta ys_{it}, \Delta ys_{it-1}]$. There are a total of 25 overidentifying moment conditions in this model.

Table 2 presents estimation results for the one-step estimator, using the weight matrix $\frac{1}{N}\sum_{i=1}^{N} Z'_i A Z_i$, and the two-step estimator.[6] The two-step estimation results are identical to those as presented in column (b) in Table 4 in Arellano and Bond (1991). Both the asymptotic standard errors and some tests based on the asymptotic variance, and the corrected versions of these are reported.

The usual asymptotic standard errors for the two-step estimator are much smaller than the standard errors for the one-step estimator. However, the standard errors that adjust for the estimation of the efficient weight matrix indicate that this perceived increase in precision is due to the downward bias of the estimates of the standard errors. The corrected standard errors are very similar to, and sometimes even larger than those of the one-step estimator. Similarly, the corrected Wald test for joint significance of the parameters is much smaller than the test based on the usual asymptotic covariance matrix.

[6] The estimation was performed using the DPD98 program for Gauss, see Arellano and Bond (1998). Standard error corrections were implemented in Gauss by the author.

Table 2. Estimation Results for Arellano-Bond (1991) data

	One-Step		Two-Step		
	coeff	std err	coeff	std err	std errc
n_{it-1}	0.5346	0.1664	0.4742	0.0853	0.1854
n_{it-2}	-0.0751	0.0680	-0.0523	0.0273	0.0517
w_{it}	-0.5916	0.1679	-0.5132	0.0493	0.1456
w_{it-1}	0.2915	0.1416	0.2246	0.0801	0.1420
k_{it}	0.3585	0.0538	0.2927	0.0395	0.0626
ys_{it}	0.5972	0.1719	0.6098	0.1085	0.1562
ys_{it-1}	-0.6117	0.2118	-0.4464	0.1248	0.2173
m_1		-2.493		-2.826	-1.999
m_2		-0.359		-0.0327	-0.316
Wald		219.6		372.0	142.0

The dependent variable is n_{it}
No. of firms 140. No. of observations 611. Time dummies included
std err are asymptotic standard errors, std errc are corrected for the estimation of the efficient weight matrix.
m_1 and m_2 are N(0,1) tests for first and second order serial correlation
Wald is a χ^2_7 test of joint significance of the coefficients

The next example uses data from Blundell and Bond (2000), who investigated estimation of production functions using the so-called system GMM estimator. A specification that was estimated is

$$\begin{aligned} (y-k)_{it} &= \alpha\,(y-k)_{it-1} + \beta\,(n-k)_{it} + \gamma\,(n-k)_{it-1} + \delta_t + u_{it} \qquad (5.1) \\ u_{it} &= \eta_i + v_{it} \end{aligned}$$

where y_{it}, n_{it} and k_{it} are the logs of sales, employment and capital stock of firm i in year t respectively. This specification accommodates first order autocorrelation in the production function and imposes constant returns to scale. The data used are a balanced panel of 509 R&D-performing US manufacturing companies observed for 8 years, 1982-1989, similar to that used in Mairesse and Hall (1996).

Table 3 presents estimation results for both the first differenced and system GMM estimators. The first differenced estimator in this case uses the $3\,(T-2)\,(T-3)\,/2$ $+\,(T-3)$ moment conditions

$$E\left(\widetilde{x}_i^{t-3}\Delta u_{it}\right) = 0 \;\; ; \;\; t = 4, ..., T$$

where $\widetilde{x}_i^{t-3} = \left(1, x_i^{t-3}\right)$, $x_i^{t-3} = (x_{i1}, ..., x_{it-3})$ and $x_{is} = (y_{is}, n_{is}, k_{is})$.

The system GMM estimator uses the $3\,(T-2)\,(T-3)\,/2$ moment conditions for the differenced equations

$$E\left(x_i^{t-3}\Delta u_{it}\right) = 0 \;\; ; \;\; t = 4, ..., T$$

plus $4\,(T-2)$ moment conditions in the levels equations

$$E\left(u_{it}\left(1, \Delta x_{it-2}\right)\right) = 0 \;\; ; \;\; t = 4, ...T.$$

The additional $3\,(T-2)$ moment conditions $E\left(u_{it}\Delta x_{it-2}\right) = 0$ are valid under certain conditions on the initial observations, see Arellano and Bover (1995) and Blundell and Bond (1998). The estimated gain in precision using the two-step GMM estimator is likely to be greater in this case, since there is no feasible one-step weight matrix that yields an asymptotically equivalent estimator to two-step GMM, even in the case of i.i.d disturbances, see Blundell and Bond (1998) for some simulation evidence.

Table 3. Estimation Results for Blundell-Bond (2000) data

	First Differences				
	One-Step		Two-Step		
	coeff	std err	coeff	std err	std errc
$(n-k)_{it}$	0.5272	0.1024	0.5731	0.0698	0.0993
$(n-k)_{it-1}$	-0.2041	0.1086	-0.1607	0.0746	0.1158
$(y-k)_{it-1}$	0.4600	0.0740	0.4146	0.0574	0.1000
m_1		-6.139		-6.210	-4.711
m_2		-0.612		-0.623	-0.583
Wald		129.5		236.02	120.1

	System				
	One-Step		Two-Step		
	coeff	std err	coeff	std err	std errc
$(n-k)_{it}$	0.5158	0.1009	0.5389	0.0598	0.0829
$(n-k)_{it-1}$	-0.2876	0.1169	-0.3155	0.0609	0.0946
$(y-k)_{it-1}$	0.5618	0.0790	0.6292	0.0371	0.0759
m_1		-6.800		-8.788	-7.737
m_2		-0.364		-0.209	-0.202
Wald		416.4		1254.7	532.5

The dependent variable is $(y-k)_{it}$
No. of firms 509. No. of observations 2545. Time dummies included
std err are asymptotic standard errors, std errc are corrected for the estimation of the efficient weight matrix.
m_1 and m_2 are N(0,1) tests for first and second order serial correlation
Wald is a χ^2_3 test of joint significance of the coefficients

The one-step estimation results presented in Table 3 are identical to those presented in columns 3 and 4 of Table 5 in Blundell and Bond (2000). For the first differenced GMM estimator there are 42 overidentifying moment conditions, whereas there are 57 overidentifying moment conditions for the system GMM estimator. Although the number of firms is quite large, again the corrected standard errors of the two-step differenced GMM estimator are much larger than the uncorrected ones. The one-step standard errors are actually smaller for two of the three coefficients. For the system two-step GMM estimator, again the corrected standard errors are larger than the uncorrected ones, but now they are smaller than the corresponding one-step standard errors. So, as expected, in this case there does appear to be a genuine gain in precision from using the efficient weight matrix.

6. Conclusions

This paper has shown that the commonly found small sample downward bias of the estimated asymptotic standard errors of the efficient two-step GMM estimator in linear models can be attributed to the fact that the asymptotic standard errors do not take account of the extra variation in small samples due to the estimated parameters in the weight matrix. A simple first order Taylor series expansion generates an extra term as a function of these initial parameter estimates, which vanishes with increasing sample size, but provides a more accurate asymptotic approximation in the case of linear moment conditions. This extra finite sample variation can be estimated and in a Monte Carlo study of a panel data model it is shown that this feasible corrected variance is close to the finite sample variance of the two-step GMM estimator.

The Monte Carlo results further show that the conventional asymptotic variance estimate of the two-step GMM estimator is a good estimate of the variance of an infeasible GMM estimator that uses the true values of the parameters to calculate the efficient weight matrix. The difference between the variances of the infeasible and feasible two-step GMM estimators can be quite large in finite samples, and the estimated corrected variance of the two-step GMM estimator captures this difference well, leading to more accurate inference.

References

[1] Arellano, M. and S.R. Bond (1991), Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, 277-98.

[2] Arellano, M. and S.R. Bond (1998), Dynamic Panel Data Estimation using DPD98 for GAUSS, http://www.ifs.org.uk/staff/steve_b.shtml.

[3] Arellano, M. and O. Bover (1995), Another Look at the Instrumental-Variable Estimation of Error-Components Models, *Journal of Econometrics*, 68, 29-52.

[4] Blundell, R.W. and S.R. Bond (1998), Initial Conditions and Moment Restrictions in Dynamic Panel Data Models, *Journal of Econometrics*, 87, 115-143.

[5] Blundell, R.W. and S.R. Bond (2000), GMM Estimation with Persistent Panel Data: An Application to Production Functions, *Econometric Reviews* 19, 321-340.

[6] Davidson, R. and J.G. MacKinnon (1996), Graphical Methods for Investigating the Size and Power of Hypothesis Tests, *Manchester School* 66, 1-26.

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

[8] Hansen, L.P., J. Heaton and A. Yaron (1996), Finite-Sample Properties of some Alternative GMM Estimators, *Journal of Business & Economic Statistics* 14, 262-280.

[9] Magnus, J.R. and H. Neudecker (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Chichester.

[10] Mairesse, J. and B.H. Hall (1996), Estimating the Productivity of Research and Development in French and US Manufacturing Firms: an Exploration of Simultaneity Issues with GMM Methods, in Wagner, K. and B. Van Ark (eds.), *International Productivity Differences and Their Explanations*, Elsevier Science, 285-315.

[11] Newey, W.K. and R.J. Smith (2000), Asymptotic Bias and Equivalence of GMM and GEL Estimators, mimeo, MIT

Appendix, The Sargan/Hansen Test for Overidentifying Restrictions

The test statistic for overidentifying restrictions in the simple linear panel data model based on the two-step GMM estimator is given by

$$S_{W_N\left(\widehat{\beta}_1\right)} = \frac{1}{N}\Delta\widehat{u}_2' Z W_N^{-1}\left(\widehat{\beta}_1\right) Z'\Delta\widehat{u}_2,$$

whereas the test statistic for overidentifying restrictions based on the infeasible GMM estimator for β, $\widehat{\beta}_{W_N(\beta_0)} = \left(X'ZW_N^{-1}(\beta_0)X'Z\right)^{-1} X'ZW_N^{-1}(\beta_0)X'y$, is given by

$$S_{W_N(\beta_0)} = \frac{1}{N}\Delta\widehat{u}_0' Z W_N^{-1}(\beta_0) Z'\Delta\widehat{u}_0,$$

where $\Delta\widehat{u}_0 = \Delta y - \widehat{\beta}_{W_N(\beta_0)}\Delta x$. Under the null that the moment conditions are valid, $S_{W_N(\beta_0)}$ and $S_{W_N\left(\widehat{\beta}_1\right)}$ both have a limiting χ^2_{q-k} distribution.

Figures 3 and 4 depicts the p-value plots for the Sargan/Hansen tests for overidentification for the two-step and infeasible GMM estimators from the Monte Carlo experiments as described in section 4, for $T = 4$ and $T = 8$ respectively. In the figures $S_{W_N\left(\widehat{\beta}_1\right)}$ is denoted SAR2 and $S_{W_N(\beta_0)}$ is denoted SAR0. It is clear that the two measures have almost exactly the same size properties, and so the size performance of the Sargan/Hansen test based on the two-step GMM estimator is not affected by the estimation of the parameters in the weight matrix. For $T = 8$, when there are many overidentifying restrictions, both test statistics are severely undersized. This is due to the fact that both $W_N(\beta_0)$ and $W_N\left(\widehat{\beta}_1\right)$ are poor estimates of the covariance of the moment conditions in this case.[7]

[7] The Sargan/Hansen test based on 3SLS estimation results, based on a weight matrix that is a

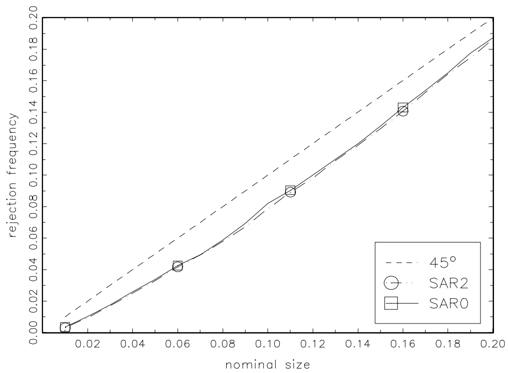

Fig 3. P-value plot, $T = 4$

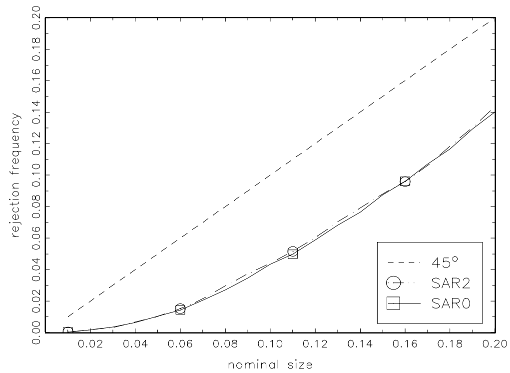

Fig 4. P-value plot, $T = 8$

The relation between $S_{W_N(\widehat{\beta}_1)}$ and $S_{W_N(\beta_0)}$ is given by

$$\begin{aligned}
S_{W_N(\widehat{\beta}_1)} &= \Delta u' Z W_N^{-1}\left(\widehat{\beta}_1\right) Z'\Delta u \\
&\quad - \Delta u' Z W_N^{-1}\left(\widehat{\beta}_1\right) Z'X\left(X'ZW_N^{-1}\left(\widehat{\beta}_1\right)Z'X\right)^{-1} X'ZW_N^{-1}\left(\widehat{\beta}_1\right) Z'\Delta u \\
&= \Delta u' Z W_N^{-1}\left(\beta_0\right) Z'\Delta u \\
&\quad - \Delta u' Z W_N^{-1}\left(\beta_0\right) Z'X\left(X'ZW_N^{-1}\left(\beta_0\right)Z'X\right)^{-1} X'ZW_N^{-1}\left(\beta_0\right) Z'\Delta u \\
&\quad + Q_{\beta_0, W_N(\beta_0)}\left(\widehat{\beta}_1 - \beta_0\right) + O_p\left(N^{-1}\right) \\
&= S_{W_N(\beta_0)} + Q_{\beta_0, W_N(\beta_0)}\left(\widehat{\beta}_1 - \beta_0\right) + O_p\left(N^{-1}\right),
\end{aligned}$$

covariance estimate that does not allow for conditional heteroscedasticity, defined as $W_N\left(\widehat{\beta}_1\right) = \frac{1}{N}\sum_{i=1}^N Z_i'\left(\frac{1}{N}\sum_{i=1}^N \Delta\widehat{u}_{1i}\Delta\widehat{u}_{1i}'\right) Z_i$, has much better size properties.

where $Q_{\beta_0,W_N(\beta_0)}$ is given by

$$
\begin{aligned}
&-u'ZW_N^{-1}(\beta_0)\frac{\partial W_N(\beta)}{\partial\beta}|_{\beta_0}W_N^{-1}(\beta_0)Z'u \\
&+2u'ZW_N^{-1}(\beta_0)\frac{\partial W_N(\beta)}{\partial\beta}|_{\beta_0}W_N^{-1}(\beta_0)Z'X\left(X'ZW_N^{-1}(\beta_0)Z'X\right)^{-1}X'ZW_N^{-1}(\beta_0)Z'u \\
&-u'ZW_N^{-1}(\beta_0)Z'X\left(X'ZW_N^{-1}(\beta_0)Z'X\right)^{-1}X'ZW_N^{-1}(\beta_0)\frac{\partial W_N(\beta)}{\partial\beta}|_{\beta_0}W_N^{-1}(\beta_0)Z'X \\
&\qquad\times\left(X'ZW_N^{-1}(\beta_0)Z'X\right)^{-1}X'ZW_N^{-1}(\beta_0)Z'u.
\end{aligned}
$$

$Q_{\beta_0,W_N(\beta_0)}\left(\widehat{\beta}_1-\beta_0\right)$ is $O_p\left(N^{-1/2}\right)$, however, the terms tend to cancel each other out. For example, in the Monte Carlo simulations, the means of the three terms of $Q_{\beta_0,W_N(\beta_0)}\left(\widehat{\beta}_1-\beta_0\right)$ are given by -0.346, 0.384 and -0.098 for $T=4$, and -1.035, 0.958 and -0.067 for $T=8$.