Division of Economics, EGC
School of Humanities and Social Sciences
Nanyang Technological University
14 Nanyang Drive
Singapore 637332

Discriminating between (in)valid external instruments and (in)valid exclusion restrictions

Jan F. Kiviet

2015

EGC Report No: 2015/08

HSS-04-86A
Tel: +65 67906073
Email: D-EGC@ntu.edu.sg

Discriminating between (in)valid external instruments and (in)valid exclusion restrictions

Jan F. Kiviet*

Version of 5 May 2016
JEL Classifications: *C01, C12, C26*
Keywords: *difference in Sargan-Hansen tests; exclusion restrictions; maintained hypothesis; instrument validity tests; overidentifying restrictions*

Abstract

In models estimated by (generalized) method of moments a test on coefficient restrictions can either be based on a Wald statistic or on the difference between evaluated criterion functions. From their correspondence it easily follows that the statistic used for testing instrument validity, the Sargan-Hansen overidentifying restrictions (OR) statistic, is equivalent to an exclusion restrictions test statistic for a nonunique group of regressor variables. We prove that asymptotically this is the case too for incremental OR tests. However, we also demonstrate that, despite this equivalence of test statistics, one can nevertheless distinguish between either the (in)validity of some additional instruments or the (un)tenability of particular exclusion restrictions. This, however, requires to be explicit about the adopted maintained hypothesis. It also highlights that recent warnings in the literature that overidentifying restrictions tests may mislead practitioners should not be directed towards the test, but to practitioners who do not realize that inference based on such tests is unavoidably conditional on the validity of particular just-identifying statistically untestable assumptions.

*Emeritus Professor of Econometrics, Amsterdam School of Economics, University of Amsterdam, PO Box 15867, 1001 NJ Amsterdam, The Netherlands (j.f.kiviet@uva.nl). Most of this paper was written while I was enjoying hospitality as Visting Professor at the Division of Economics, School of Humanities and Social Sciences, Nanyang Technological University, 14 Nanyang Drive, Singapore 637332. Helpful comments by João Santos Silva on an earlier version of this paper and further suggestions by JEM co-editor Jason Abrevaya are gratefully acknowledged.

1. Introduction

In empirical econometric models it is often not unlikely that some explanatory variables are in fact endogenous, because their realizations are contemporaneously correlated with the error terms of the model. To overcome inference problems due to such correlation one often exploits external (non-explanatory) variables that should be uncorrelated with these errors. To verify whether these variables (instruments) seem uncorrelated with the disturbances indeed one can use a so-called test for overidentifying restrictions (OR), see Sargan (1958), Hansen (1982). The mandatory use of OR tests on a routine basis has been advocated at various places in the literature. See, for instance, the list mentioned in Baum et al. (2003, footnote 11). However, warnings that OR tests may mislead practitioners were recently issued in Parente and Santos Silva (2012) and Guggenberger (2012), and can also be found in, for instance, Hayashi (2000, p.218) and Cameron and Trivedi (2005, p.277).

In this study we will place these warnings regarding OR tests into perspective and highlight that they only apply if one is hesitant with respect to formulating an explicit maintained hypothesis, involving sufficient a priori orthogonality conditions and exclusion restrictions. Also we will expose the algebraic equivalences between OR and CR (coefficient restrictions) test statistics and explain why these tests nevertheless allow distinct inferences. Moreover, we discuss to what degree similar issues arise for incremental OR tests.

For the sake of simplicity we will focus here mainly on the linear method of moments context, but our derivations can be extended to more general settings, as we shall indicate. First, while introducing our notation, we review some general results on testing in a linear instrumental variables context, such as illustrating that the classical trinity of test principles established in the context of Maximum Likelihood inference, constituted by Wald (W), Likelihood Ratio (LR) and Lagrange Multiplier (LM), have their counterparts in a semiparametric setting in which models can efficiently and robustly be estimated by (generalized) method of moments, see also Newey and West (1987). In a linear method of moments context CR tests based on a W statistic are in fact equivalent to the test obtained by taking the difference between evaluated criterion functions (CF) for the restricted and the unrestricted model. The latter reminiscences of an LR-type test. By employing an estimate for the disturbance variance obtained from the restricted model both W and CF can be implemented as a kind of LM-type test. A very particular implementation of a CF test gives the Sargan-Hansen statistic for testing overidentifying restrictions. Thus, the very same test statistic can be interpreted either as a CR test or as an OR instrument validity test. However, we will highlight that these two interpretations require different maintained hypotheses. This has not always been made

very clear in earlier literature.

We also prove that in linear models similar equivalences and differences exist between exclusion restrictions tests and so-called incremental overidentifying restrictions (IOR) tests (also addressed as difference in Sargan-Hansen tests), which verify the validity of a subset of instruments. From these findings we conclude that strictly following formal statistical test principles enables to characterize the context in which OR and IOR tests should not mislead and in fact allow to distinguish inference regarding the (in)validity of some additional orthogonality conditions from inference on the (un)tenability of particular exclusion restrictions. Practical problems do emerge only when one is unable or unwilling to condition the analysis on a firm initial maintained hypothesis.

The structure of this study is as follows. In Section 2 we introduce the model and various test procedures and concepts. Section 3 demonstrates how the standard Sargan test can either be used for testing instrument validity or under a different maintained hypothesis can be interpreted as a CR test. Section 4 demonstrates to what degree these results do apply to incremental Sargan tests as well. Finally Section 5 indicates options for further generalizations and summarizes the conclusions.

2. Corresponding test principles

We focus on the single linear simultaneous regression model $y = X\beta + \varepsilon$ with $\varepsilon \sim (0, \sigma_\varepsilon^2 I)$, where $X = (x_1, ..., x_n)'$ is an $n \times K$ matrix of rank K with unknown coefficient vector $\beta \in \mathbb{R}^K$. To cope with $E(x_i\varepsilon_i) = \sigma_{x\varepsilon} \neq 0$ for $i = 1, ..., n$ we will employ an $n \times L$ instrumental variables matrix $Z = (z_1, ..., z_n)'$ of rank $L \geq K$. The instrumental variables (IV) or two-stage least-squares (2SLS) estimator is given by

$$\hat{\beta} = (X'P_Z X)^{-1} X'P_Z y, \tag{2.1}$$

where $P_A = A(A'A)^{-1}A'$ for any full column rank matrix A. When $E(z_i\varepsilon_i) = 0$ (the instruments are valid) and standard regularity conditions are fulfilled too (including sufficient relevance of the instruments) then the IV estimator is consistent and asymptotically normal, whereas its variance can be estimated by $\widehat{Var}(\hat{\beta}) = \hat{\sigma}_\varepsilon^2 (X'P_Z X)^{-1}$, where $\hat{\sigma}_\varepsilon^2 = \hat{\varepsilon}'\hat{\varepsilon}/n$ with residuals $\hat{\varepsilon} = y - X\hat{\beta}$.

A test on any set of $K_2 < K$ linear restrictions on the coefficients β can after a simple transformation of the model be represented as testing exclusion restrictions $\mathcal{H}_0 : \beta_2 = 0$ in

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \text{ with } \varepsilon \sim (0, \sigma_\varepsilon^2 I), \tag{2.2}$$

where X_j is $n \times K_j$ and β_j is $K_j \times 1$ for $j = 1, 2$. Making use of standard results on partitioned regression applied to the second-stage regression, in which y is regressed on

$P_Z X = (P_Z X_1, P_Z X_2) = (\hat{X}_1, \hat{X}_2)$, one easily finds that $\hat{\beta} = (\hat{\beta}_1', \hat{\beta}_2')'$, with

$$\hat{\beta}_2 = (\hat{X}_2' M_{\hat{X}_1} \hat{X}_2)^{-1} \hat{X}_2' M_{\hat{X}_1} y \tag{2.3}$$

and $\widehat{Var}(\hat{\beta}_2) = \hat{\sigma}_\varepsilon^2 (\hat{X}_2' M_{\hat{X}_1} \hat{X}_2)^{-1}$, where $M_A = I - P_A$. The W statistic for testing $\mathcal{H}_0 : \beta_2 = 0$ against alternative $\mathcal{H}_1 : \beta_2 \neq 0$ readily follows and is given by

$$W_{\beta_2} = \hat{\beta}_2' [\widehat{Var}(\hat{\beta}_2)]^{-1} \hat{\beta}_2. \tag{2.4}$$

Under $\mathcal{H}_0$ it is asymptotically $\chi^2(K_2)$ distributed, provided the maintained hypothesis does hold indeed. The maintained hypothesis (which is the union of $\mathcal{H}_0$ and $\mathcal{H}_1$ and all underlying assumptions) can here be characterized loosely as

$\mathcal{M}_1$*: (i) the actual data generating process (DGP) is nested in model (2.2) with* K regressors*; (ii)* X *and* Z *show sufficient regularity; (iii)* $n^{-1/2} Z'\varepsilon \xrightarrow{d} \mathcal{N}(0, \sigma_\varepsilon^2 \operatorname{plim} n^{-1} Z'Z)$.

Substituting (2.3) in (2.4) and making use of a result on projection matrices for a partitioned full column rank matrix $A = (A_1, A_2)$ which says $P_A = P_{A_1} + P_{M_{A_1} A_2}$, we find

$$W_{\beta_2} = y' P_{M_{\hat{X}_1} \hat{X}_2} y / \hat{\sigma}_\varepsilon^2 = y' (P_{\hat{X}} - P_{\hat{X}_1}) y / \hat{\sigma}_\varepsilon^2 = y' (M_{\hat{X}_1} - M_{\hat{X}}) y / \hat{\sigma}_\varepsilon^2. \tag{2.5}$$

Hence, the test statistic can easily be obtained by taking the difference between the restricted and the unrestricted sum of squared residuals of second stage regressions, while scaling by a consistent estimator of σ_ε^2. If the latter is obtained from the restricted residuals $\tilde{\varepsilon} = y - X_1 \tilde{\beta}_1$, where

$$\tilde{\beta}_1 = (\hat{X}_1' \hat{X}_1)^{-1} \hat{X}_1' y \tag{2.6}$$

is the restricted IV estimator (imposing $\beta_2 = 0$) and $\tilde{\sigma}_\varepsilon^2 = \tilde{\varepsilon}'\tilde{\varepsilon}/n$, the test statistic reminds of a Lagrange Multiplier test. Therefore we indicate it as

$$LM_{\beta_2} = y' (M_{\hat{X}_1} - M_{\hat{X}}) y / \tilde{\sigma}_\varepsilon^2, \tag{2.7}$$

which is asymptotically equivalent with W_{β_2}.

A test statistic with similarities to the LR principle is found as follows. Estimator $\hat{\beta}$ is obtained by minimizing with respect to β a quadratic form in the vector $Z'(y - X\beta)$ in which a weighting matrix is used proportional to $(Z'Z)^{-1}$. This yields a consistent estimator with optimal variance under the maintained hypothesis. Defining the criterion function as

$$Q(\beta; y, X, Z, \sigma_\varepsilon^2) = (y - X\beta)' Z (Z'Z)^{-1} Z' (y - X\beta) / \sigma_\varepsilon^2, \tag{2.8}$$

one can verify that it attains its minimum for $\hat{\beta}$ of (2.1), giving

$$\begin{aligned} Q(\hat{\beta}; y, X, Z, \sigma_\varepsilon^2) &= (y - X\hat{\beta})' P_Z (y - X\hat{\beta}) / \sigma_\varepsilon^2 = (y' P_Z y - 2\hat{\beta}' \hat{X} y + \hat{\beta}' \hat{X}' \hat{X} \hat{\beta}) / \sigma_\varepsilon^2 \\ &= (y' P_Z y - y' P_{\hat{X}} y) / \sigma_\varepsilon^2 = (y' M_{\hat{X}} y - y' M_Z y) / \sigma_\varepsilon^2, \end{aligned}$$

which has in its numerator the difference between the sum of the squares of the second-stage residuals and the reduced form residuals. In the model under $\beta_2 = 0$, upon using the criterion function $Q(\beta_1; y, X_1, Z, \sigma_\varepsilon^2)$, one finds $\tilde{\beta}_1$ of (2.6) with $Q(\tilde{\beta}_1; y, X_1, Z, \sigma_\varepsilon^2) = (y'M_{\hat{X}_1}y - y'M_Z y)/\sigma_\varepsilon^2$. The difference between these two minimized criterion functions is

$$Q(\tilde{\beta}_1; y, X_1, Z, \sigma_\varepsilon^2) - Q(\hat{\beta}; y, X, Z, \sigma_\varepsilon^2) = (y'M_{\hat{X}_1}y - y'M_{\hat{X}}y)/\sigma_\varepsilon^2.$$

Now substituting either $\hat{\sigma}_\varepsilon^2$ or $\tilde{\sigma}_\varepsilon^2$, we find that taking the difference between these two evaluated criterion functions yields either

$$CF_{\beta_2}(\hat{\sigma}_\varepsilon^2) = Q(\tilde{\beta}_1; y, X_1, Z, \hat{\sigma}_\varepsilon^2) - Q(\hat{\beta}; y, X, Z, \hat{\sigma}_\varepsilon^2) = W_{\beta_2} \tag{2.9}$$

or

$$CF_{\beta_2}(\tilde{\sigma}_\varepsilon^2) = LM_{\beta_2}. \tag{2.10}$$

These equivalences are algebraic, thus hold irrespective of the validity of $\mathcal{M}_1$.

3. The Sargan test and exclusion restrictions

A special situation occurs in case $L = K$. This specializes $\mathcal{M}_1$ to what we will indicate by $\mathcal{M}_1^{L=K}$. Under this maintained hypothesis the model is just identified, but overidentified under the K_2 coefficient restrictions $\beta_2 = 0$. When the model is just identified then, when minimizing $Q(\beta; y, X, Z, \sigma_\varepsilon^2)$, the $L = K$ orthogonality conditions $Z'(y - X\hat{\beta}) = 0$ will all be satisfied in the sample, giving $Q(\hat{\beta}; y, X, Z, \sigma_\varepsilon^2) = 0$. So, in this particular situation, (2.9) specializes into

$$W_{\beta_2}^{L=K} = CF_{\beta_2}^{L=K}(\hat{\sigma}_\varepsilon^2) = Q(\tilde{\beta}_1; y, X_1, Z, \hat{\sigma}_\varepsilon^2) = \tilde{\varepsilon}'P_Z\tilde{\varepsilon}/\hat{\sigma}_\varepsilon^2, \tag{3.1}$$

and (2.10) into the asymptotically equivalent statistic

$$LM_{\beta_2}^{L=K} = CF_{\beta_2}^{L=K}(\tilde{\sigma}_\varepsilon^2) = \tilde{\varepsilon}'P_Z\tilde{\varepsilon}/\tilde{\sigma}_\varepsilon^2. \tag{3.2}$$

The latter expression is the well-known Sargan test statistic which is used to check the validity of the instruments by testing the $L-K_1$ overidentifying restrictions in the model which has just the K_1 regressors X_1, and can be expressed as

$$y = X_1\beta_1^* + \varepsilon^* \text{ with } \varepsilon^* \sim (0, \sigma_{\varepsilon^*}^2 I). \tag{3.3}$$

From its algebraic equivalence with a CR test it is immediately obvious that testing for OR has similarities with testing the validity of $L - K_1$ exclusion restrictions.

However, there are dissimilarities as well. Validity of the L orthogonality conditions $E(z_i\varepsilon_i) = 0$ is part of the maintained hypothesis $\mathcal{M}_1^{L=K}$ for the test of $L - K_1$ exclusion

restrictions $\beta_2 = 0$ in model (2.2), whereas the Sargan test is used to verify validity of the orthogonality conditions $E(z_i\varepsilon_i^*) = 0$ in model (3.3). Although using exactly the same test statistic and critical values, we will explicate that the two tests have intrinsically different $\mathcal{H}_0$ and $\mathcal{H}_1$ and build on different maintained hypotheses. We are not aware of literature where these differences and their consequences have been spelled out in full detail. Below we aim to fill this gap.

Note that expression (3.2) is invariant regarding X_2. So, replacing X_2 by any alternative $n \times K_2$ matrix X_2^a would yield equivalent results for $CF_{\beta_2}^{L=K}(\tilde{\sigma}_\varepsilon^2)$, provided $P_Z(X_1, X_2^a)$ has full column rank $K = L$. Thus, (X_1, X_2^a) should have full column rank and the variables X_2^a should be sufficiently related to the instruments Z, but at the same time they may also be endogenous regarding ε. Due to the projection of the regressors on the space spanned by the instruments Z in the second stage regression, and the presence of the regressors X_1, the CR test actually does not test the explanatory power of X_2^a as such, but just that of $M_{\hat{X}_1}P_Z X_2^a = (P_Z - P_{\hat{X}_1}P_Z)X_2^a = (P_Z - P_{\hat{X}_1})X_2^a$, which is its projection on the space spanned by Z, as far as this is orthogonal to $P_Z X_1$.

A viable choice for X_2^a would be the following. Let $F = (F_1, F_2)$ be a full rank $L \times L$ matrix, where F_1 is $L \times L_1$, F_2 is $L \times L_2$, with $L = L_1 + L_2$ and $L_1 = K_1$. Now consider the transformed instruments $Z^f = (Z_1^f, Z_2^f) = ZF = (ZF_1, ZF_2)$, and let F_1 be such that square matrix $Z_1^{f\prime}X_1$ has full rank. Obviously, F is nonunique. Note that if the transformed instruments Z_1^f are not just relevant for X_1, but were valid for the errors ε^* of model (3.3) as well, then the coefficients β_1^* would be just identified by them. When taking $X_2^a = M_{Z_1^f}Z_2^f$, then $P_Z(X_1, X_2^a)$ has full column rank as required, and statistic $CF_{\beta_2^{**}}^{L=K}(\tilde{\sigma}_\varepsilon^2)$ for testing $\mathcal{H}_0 : \beta_2^{**} = 0$ in model

$$y = X_1\beta_1^{**} + M_{Z_1^f}Z_2^f\beta_2^{**} + \varepsilon^{**} \tag{3.4}$$

is therefore equivalent to the Sargan test statistic $CF_{\beta_2}^{L=K}(\tilde{\sigma}_\varepsilon^2)$. Estimating β_1^{**} of (3.4) by IV using the instruments Z or Z^f (which is equivalent because $P_Z = P_{Z^f}$) yields

$$\begin{aligned}\hat{\beta}_1^{**} &= (X_1'P_{Z^f}M_{M_{Z_1^f}Z_2^f}P_{Z^f}X_1)^{-1}X_1'P_{Z^f}M_{M_{Z_1^f}Z_2^f}P_{Z^f}y \\ &= (X_1'P_{Z_1^f}X_1)^{-1}X_1'P_{Z_1^f}y = (Z_1^{f\prime}X_1)^{-1}Z_1^{f\prime}y = \hat{\beta}_1^*,\end{aligned} \tag{3.5}$$

because $P_{Z^f}M_{M_{Z_1^f}Z_2^f}P_{Z^f} = P_{Z^f} - P_{M_{Z_1^f}Z_2^f} = P_{Z_1^f}$. Hence, augmenting model (3.3) by K_2 regressors $M_{Z_1^f}Z_2^f$ and using instruments Z or Z^f yields coefficient estimates $\hat{\beta}_1^{**}$ for the regressors X_1 which are equivalent to the simple IV estimator $\hat{\beta}_1^*$ obtained in model (3.3) when just using the K_1 instruments Z_1^f, irrespective of the validity of any of the instruments.

Apart from the interpretation of $CF_{\beta_2}^{L=K}(\tilde{\sigma}_\varepsilon^2)$ as testing $K_2 = L - K_1$ exclusion restrictions $\beta_2 = 0$ in model (2.2) under maintained hypothesis $\mathcal{M}_1^{L=K}$, which presupposes

the K orthogonality conditions $E(z_i\varepsilon_i) = 0$, yet another coherent interpretation of the numerically equivalent test statistic $CF^{L=K}_{\beta_2^{**}}(\tilde{\sigma}^2_\varepsilon)$ is the following. It tests whether the $L - K_1$ variables Z_2^f are additional valid external instruments for model (3.3) under the maintained hypothesis given by

$\mathcal{M}_2^{L=K}$*: (i) the DGP is nested in model (3.3) which has* $K_1 < K = L$ regressors*; (ii)* X_1 *and* Z *(and thus* Z^f*) show sufficient regularity; (iii)* $n^{-1/2}Z_1^{f\prime}\varepsilon^* \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\varepsilon^*} \operatorname{plim} n^{-1}Z_1^{f\prime}Z_1^f)$.

Note that for adopting $\mathcal{M}_2^{L=K}$ in practice it might be desirable to specify matrix F_1; this issue will be addressed below.

Under $\mathcal{M}_2^{L=K}$ the value of $E(z_{2i}^f\varepsilon_i^*) = \sigma_{z_2^f\varepsilon^*}$ is unspecified. Obviously, if next to Z_1^f the variables Z_2^f are uncorrelated with ε^* then $M_{Z_1^f}Z_2^f$ should not have explanatory power in model (3.4). Hence, when adopting $\mathcal{M}_2^{L=K}$, then under the null hypothesis $\beta^{**} = 0$ in model (3.4) all instruments Z are valid for model (3.3), implying $\sigma_{z_2^f\varepsilon^*} = 0$ so that $\tilde{\beta}_1$ is consistent, asymptotically normal and efficient for β_1. Below we will simply use the acronym AE (asymptotically efficient) to indicate for an estimator if, under the assumptions made, it is consistent with asymptotically normal distribution and smallest asymptotic variance (in a matrix sense). The alternative $\beta^{**} \neq 0$ implies $\sigma_{z_2^f\varepsilon^*} \neq 0$. So, in the context of model (3.4) test statistic $CF^{L=K}_{\beta_2^{**}}(\tilde{\sigma}^2_\varepsilon)$ tests the null hypothesis given by the $L - K_1$ orthogonality conditions $E(z_{2i}^f\varepsilon_i^*) = 0$ or overidentifying restrictions $\beta^{**} = 0$. When these conditions/restrictions are invalid then, given validity of $\mathcal{M}_2^{L=K}$, estimator $\hat{\beta}_1^* = (Z_1^{f\prime}X_1)^{-1}Z_1^{f\prime}y$ for β_1^* is AE in model (3.3). In the extended regression model (3.4) all variables in both Z_1^f and Z_2^f constitute valid instruments regarding ε^{**} under $\mathcal{M}_2^{L=K}$, because inclusion of the regressors $M_{Z_1^f}Z_2^f$ purges the disturbances from their correlation with Z_2^f. That this regression too produces the AE estimator $\hat{\beta}_1^*$ for the coefficients β_1, as is proved by (3.5), is because it constitutes a special case of the so-called "partialling out" result which will be reviewed at the end of the next section.

It is the case that truth of the null $\beta_2 = 0$ under $\mathcal{M}_1^{L=K}$ implies $E(z_i\varepsilon_i^*) = 0$, and so does truth of the null $E(z_{2i}^f\varepsilon_i^*) = 0$ under $\mathcal{M}_2^{L=K}$. Under both null hypotheses estimator $\tilde{\beta}_1$ is AE for β_1. However, imposing the respective null hypotheses implies either accepting coefficient restrictions or accepting extra orthogonality assumptions. Invalidity of $\beta_2 = 0$ under $\mathcal{M}_1^{L=K}$ renders $\hat{\beta}$ the AE estimator for β, but invalidity of $E(z_{2i}^*\varepsilon_i^*) = 0$ under $\mathcal{M}_2^{L=K}$ renders $\hat{\beta}_1^*$ AE for the coefficients of just the regressors X_1, since in this setup the DGP is supposed to be nested in model (3.3) which has just K_1 regressors.

So, on the basis of the single test statistic $LM^{L=K}_{\beta_2}(\tilde{\sigma}^2_\varepsilon) = CF^{L=K}_{\beta_2^{**}}(\tilde{\sigma}^2_\varepsilon)$, or the asymptotically equivalent $W^{L=K}_{\beta_2}(\hat{\sigma}^2_\varepsilon)$, one can discriminate between either (in)validity of $L-K_1$ exclusion restrictions in model (2.2) or (in)validity of $L-K_1$ external instruments in model (3.3). This just hinges upon which of the nonnested maintained hypotheses has actually been adopted: $\mathcal{M}_1^{L=K}$ or $\mathcal{M}_2^{L=K}$. In both contexts the tests concern a possible

loss/gain in the degree of overidentification by $L - K_1$. Namely, either due to rejecting/accepting exclusion restrictions on the coefficients of potential explanatory variables X_2 in model (2.2), or on accepting/rejecting exclusion restrictions on the coefficients of the auxiliary regressors $M_{Z_1^f} Z_2^f$ in model (3.4). The latter implies accepting/rejecting validity of $L - K_1$ external instruments Z_2^f.

The warnings of Parente and Santos Silva (2012) and Gugenberger (2012) concern the problems that emerge when using the Sargan statistic for testing jointly the L orthogonality conditions $E(z_i \varepsilon_i^*) = 0$ in model (3.3). However, using this L dimensional null hypothesis, is asking too much from this $L - K_1$ degrees of freedom test statistic, leading to inconsistency of the test, meaning: existence of nonlocal alternatives to the null that are not rejected with probability one asymptotically. The classic Sargan test on instrument validity is only consistent in case one is willing to adopt maintained hypothesis $\mathcal{M}_2^{L=K}$. It is uncomfortable but also unavoidable that this maintained hypothesis embraces the statistically untestable orthogonality conditions expressed by $E(z_{1i}^f \varepsilon_i^*) = 0$ $(i = 1, ..., n)$. These untestable conditions entail: It is possible to find K_1 linearly independent linear combinations of the K variables in matrix Z which establish valid instruments for model (3.3) and ensure just identification of β_1^*.

It is the researcher who may decide whether (s)he leaves these K_1 orthogonality conditions implicit, or makes them explicit by actually specifying the matrix F_1, for instance by choosing $F_1 = (I, O)'$ and then thus simply assuming that the first K_1 instruments in the matrix Z are valid in model (3.3). In that special case the explicit null hypothesis of the Sargan test is that the instruments $M_{Z_1} Z_2$ are valid too, which would imply $E(z_{i2} \varepsilon_i^*) = 0$ in model (3.3).

4. The incremental Sargan test and exclusion restrictions

The validity of a subset of the instruments Z_2^f for model (3.3) can be tested by a so-called incremental Sargan or IOR test. For that we will examine too in what way such a procedure has (dis)similarities with testing exclusion restrictions. However, for this analysis we will take as our starting point the usual model with K regressors and $L > K$ instruments and therefore have to introduce some further notation.

We consider a partition of the matrix of L instruments, denoted as $Z = (Z_m, Z_a)$, where for $j \in \{m, a\}$ matrix $Z_j = (z_{j1}, ..., z_{jn})'$ is $n \times L_j$. Matrix Z_m, with $L_m \geq K$, contains the instruments which are maintained to be valid, whereas the validity of the additional $L_a > 0$ instruments Z_a will be examined by testing. So the maintained hypothesis is now given by

$\mathcal{M}_2^{L>K}$: *(i) the DGP is nested in model* $y = X\beta + \varepsilon$, *which has* K *regressors; (ii)* X

and $Z = (Z_m, Z_a)$ *with* $L > L_m \geq K$ *show sufficient regularity; (iii)* $n^{-1/2} Z_m' \varepsilon \xrightarrow{d} \mathcal{N}(0, \sigma_\varepsilon^2 \operatorname{plim} n^{-1} Z_m' Z_m)$.

Just using the instruments Z_m yields estimator and residuals

$$\hat{\beta}_m = (X' P_{Z_m} X)^{-1} X' P_{Z_m} y, \ \hat{\varepsilon}_m = y - X\hat{\beta}_m. \tag{4.1}$$

From the foregoing section it follows that the Sargan statistic

$$S_m = n \cdot \hat{\varepsilon}_m' P_{Z_m} \hat{\varepsilon}_m / \hat{\varepsilon}_m' \hat{\varepsilon}_m \tag{4.2}$$

is asymptotically $\chi^2(L_m - K)$ distributed under $\mathcal{M}_2^{L>K}$ when $L_m > K$, whereas it is zero for $L_m = K$. Using the additional instruments as well yields $\hat{\beta}$ given in (2.1) and $\hat{\varepsilon} = y - X\hat{\beta}$ and enables to calculate

$$S_{ma} = n \cdot \hat{\varepsilon}' P_Z \hat{\varepsilon} / \hat{\varepsilon}' \hat{\varepsilon}. \tag{4.3}$$

Now the null hypothesis $E(z_{ia}\varepsilon_i) = 0$, which involves L_a orthogonality conditions for model (2.2) additional to the L_m orthogonality conditions already implied by $\mathcal{M}_2^{L>K}$, can be tested by the incremental Sargan statistic

$$IS_a = S_{ma} - S_m, \tag{4.4}$$

which under $\mathcal{M}_2^{L>K}$ has asymptotic null distribution $\chi^2(L_a)$. Note that for $L_m = K$ statistic IS_a corresponds to the Sargan test examined in the foregoing section, but now the role of X_1 is plaid by X, so K_1 is now K, whereas K_2 is given here by L_a and what earlier was called $\tilde{\varepsilon}$ is given by $\hat{\varepsilon}$ now.

Most published derivations of the null distribution of IS_a start off from a GMM (generalized method of moments) context. Our simple linear framework allows a very concise derivation, which goes as follows. Without loss of generality we assume $Z_m' Z_a = O$, which may have been achieved by replacing the original Z_a by $M_{Z_m} Z_a$. P_Z is not affected by this partial orthogonalization, but now $P_Z = P_{Z_m} + P_{Z_a}$. From $P_Z \hat{\varepsilon} = P_Z[I - X(X' P_Z X)^{-1} X' P_Z]\varepsilon = (P_Z - P_{P_Z X})\varepsilon$ we find $\hat{\varepsilon}' P_Z \hat{\varepsilon} = \varepsilon' P_Z \varepsilon - \varepsilon' P_{P_Z X} \varepsilon = \varepsilon'(P_{Z_m} + P_{Z_a} - P_{P_Z X})\varepsilon$ and $\hat{\varepsilon}_m' P_{Z_m} \hat{\varepsilon}_m = \varepsilon'(P_{Z_m} - P_{P_{Z_m} X})\varepsilon$, so $\hat{\varepsilon}' P_Z \hat{\varepsilon} - \hat{\varepsilon}_m' P_{Z_m} \hat{\varepsilon}_m = \varepsilon'(P_{Z_a} + P_{P_{Z_m} X} - P_{P_Z X})\varepsilon$. The matrix in this quadratic form is symmetric and idempotent. The latter follows from $P_{Z_a} P_{P_{Z_m} X} = O$ and $(P_{Z_a} + P_{P_{Z_m} X}) P_{P_Z X} = (P_{Z_a} + P_{Z_m}) X (X' P_Z Z)^{-1} X' P_Z = P_{P_Z X}$. Since $tr(P_{Z_a} + P_{P_{Z_m} X} - P_{P_Z X}) = L_a + K - K = L_a$ this means that under $\mathcal{M}_2^{L>K}$ and the null hypothesis $E(z_{ia}\varepsilon_i) = 0$ $(i = 1, ..., n)$, so when all instrument in Z are valid, $IS_a \xrightarrow{d} \varepsilon'(P_{Z_a} + P_{P_{Z_m} X} - P_{P_Z X})\varepsilon / \sigma_\varepsilon^2 \xrightarrow{d} \chi^2(L_a)$, because under the null both $\operatorname{plim} \hat{\varepsilon}_m' \hat{\varepsilon}_m / n = \sigma_\varepsilon^2$ and $\operatorname{plim} \hat{\varepsilon}' \hat{\varepsilon} / n = \sigma_\varepsilon^2$.

We shall now examine tests for omitted regressors (or exclusion restrictions) and seek cases which show correspondences with statistic IS_a. Consider the extended model with $K_a \leq L - K$ additional regressors X_a given by

$$y = X\beta + X_a\beta_a + \varepsilon_a, \text{ with } \varepsilon_a \sim (0, \sigma^2_{\varepsilon_a} I), \tag{4.5}$$

and the maintained hypothesis defined by

$\mathcal{M}_1^{L>K}$: *(i) the DGP is nested in model* $y = X\beta + X_a\beta_a + \varepsilon_a$, *which has* $K + K_a \leq L$ *regressors; (ii)* (X, X_a) *and* Z *show sufficient regularity; (iii)* $n^{-1/2}Z'\varepsilon_a \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\varepsilon_a} \operatorname{plim} n^{-1}Z'Z)$.

Let us indicate the IV results when estimating model (4.5) using instruments Z as $y = X\hat{\beta}_* + X_a\hat{\beta}_a + \hat{\varepsilon}_a$. Then the LM-like test statistic for $\mathcal{H}_0 : \beta_a = 0$ is given by

$$CF_{\beta_a}(\hat{\sigma}^2_\varepsilon) = (\hat{\varepsilon}'P_Z\hat{\varepsilon} - \hat{\varepsilon}'_a P_Z \hat{\varepsilon}_a)/\hat{\sigma}^2_\varepsilon, \tag{4.6}$$

which under $\mathcal{M}_1^{L>K}$ has asymptotic null distribution $\chi^2(K_a)$.

Still assuming without loss of generality that $Z'_m Z_a = O$, we shall first examine the specific case $X_a = Z_a$, for which expression (4.6) can be specialized by using the following results. Substituting $X_a = Z_a$ we obtain

$$\hat{\beta}_* = (X'P_Z M_{Z_a} P_Z X)^{-1} X'P_Z M_{Z_a} y = (X'P_{Z_m}X)^{-1}X'P_{Z_m}y = \hat{\beta}_m, \tag{4.7}$$

because $M_{Z_a}P_Z = P_Z - P_{Z_a}P_Z = P_Z - P_{Z_a} = P_{Z_m}$. And, since in IV estimation the residuals are orthogonal to the second-stage regressors, we have $Z'_a\hat{\varepsilon}_a = 0$. This yields $P_Z\hat{\varepsilon}_a = (P_{Z_m} + P_{Z_a})\hat{\varepsilon}_a = P_{Z_m}(y - X\hat{\beta}_* - Z_a\hat{\beta}_a) = P_{Z_m}(y - X\hat{\beta}_*) = P_{Z_m}(y - X\hat{\beta}_m) = P_{Z_m}\hat{\varepsilon}_m$, where we used (4.7) for the last equality. From this we find

$$\begin{aligned} CF^{X_a=Z_a}_{\beta_a}(\hat{\sigma}^2_\varepsilon) &= (\hat{\varepsilon}'P_Z\hat{\varepsilon} - \hat{\varepsilon}'_a P_Z\hat{\varepsilon}_a)/\hat{\sigma}^2_\varepsilon = (\hat{\varepsilon}'P_Z\hat{\varepsilon} - \hat{\varepsilon}'_m P_{Z_m}\hat{\varepsilon}_m)/\hat{\sigma}^2_\varepsilon \\ &= S_{ma} - S_m \frac{\hat{\varepsilon}'_m\hat{\varepsilon}_m}{\hat{\varepsilon}'\hat{\varepsilon}}. \end{aligned} \tag{4.8}$$

Under $\mathcal{M}_1^{L>K}$ and $\mathcal{H}_0 : \beta_a = 0$ both $\hat{\varepsilon}'_m\hat{\varepsilon}_m/n$ and $\hat{\varepsilon}'\hat{\varepsilon}/n$ converge to σ^2_ε, thus the ratio $\hat{\varepsilon}'_m\hat{\varepsilon}_m/\hat{\varepsilon}'\hat{\varepsilon}$ equals 1 asymptotically. Therefore, under their respective null hypotheses, the incremental Sargan test statistic $IS_a = S_{ma} - S_m$ is asymptotically equivalent with the exclusion restrictions test statistic $CF^{X_a=Z_a}_{\beta^*_a}(\hat{\sigma}^2_\varepsilon)$.

In fact, this also holds for more general matrices than just $X_a = Z_a$. Consider instead

$$X^*_a = Z_aC_1 + XC_2 + M_ZC_3 + \varepsilon C_4, \tag{4.9}$$

where the finite matrices C_1, C_2, C_3, C_4 are $L_a \times K_a$, $K \times K_a$, $n \times K_a$ and $1 \times K_a$ respectively. Now writing $\hat{X}^*_a = P_Z X^*_a$ and $\hat{X} = P_Z X$ it follows that $\hat{X}^*_a = Z_aC_1 + \hat{X}C_2 + P_Z\varepsilon C_4$. Under the null hypothesis $\beta_a = 0$, we have $\operatorname{plim} n^{-1}Z'\varepsilon = 0$, hence

$P_Z \varepsilon C_4 \to O$, thus $M_{\hat{X}} \hat{X}_a^* \to M_{\hat{X}} Z_a C_1$. The regularity assumption of $\mathcal{M}_1^{L>K}$ requires this to have full column rank, hence C_1 should have rank $K_a \leq L_a$. If $K_a = L_a$ and C_1 has full rank we obtain $P_{M_{\hat{X}} \hat{X}_a^*} \to P_{M_{\hat{X}} Z_a}$. Thus, tests for the omission of regressors X_a which belong to group (4.9) with C_1 of full rank $K_a = L_a$ will yield test statistics which under the null are all asymptotically equivalent with statistic IS_a.

Estimator $\hat{\beta}$ is AE both under $\mathcal{M}_1^{L>K}$ with valid exclusion restrictions $\beta_a = 0$ and under $\mathcal{M}_2^{L>K}$ with valid additional instruments Z_a. However, under $\mathcal{M}_2^{L>K}$ while instruments Z_a are invalid $\hat{\beta}_m$ is AE, whereas under $\mathcal{M}_1^{L>K}$ with $\beta_a \neq 0$ the estimators $\hat{\beta}_*$ and $\hat{\beta}_a$ are AE in model (4.5), yielding an estimator for β_* asymptotically equivalent to $\hat{\beta}_m$ only if X_a belongs to group (4.9) with $K_a = L_a$, C_1 of full rank and $C_2 = O$.

The algebraic estimator equivalence for $X_a = Z_a$ given in (4.7) reestablishes the so-called partialling out result, which says that IV coefficient estimates are invariant when (putative) exogenous regressors are deliberately omitted, provided they are also removed from and filtered out of the remaining instruments. In Hendry (2011) this analytic result is established by simulation, which leads to the unsatisfactory conclusion that (in our notation) $\hat{\beta}_*$ and $\hat{\beta}_m$ "hardly differ". Moreover, in the simulation design used the instruments are all valid, $L = 3$, $K = 2$, $L_a = 1$ and all variables are normally distributed, whereas neither of these is required for the algebraic result to hold exactly.

5. Interpretation and conclusion

The above results lead to clear guidelines only when a researcher is willing to adopt unambiguously a particular maintained hypothesis, which (s)he should according to sound Neyman-Pearson statistical test methodology. In practice, however, a more flexible though opportunistic approach may often seem more appealing. The two juxtaposed maintained hypotheses $\mathcal{M}_1^{L=K}$ and $\mathcal{M}_2^{L=K}$ in Section 3, and their generalizations $\mathcal{M}_1^{L>K}$ and $\mathcal{M}_2^{L>K}$ in Section 4, both suppose that knowledge is available already of a model specification in which the DGP is nested, whereas endeavors to reach that stage form usually the major challenge in an empirical modelling exercise. This explains why researchers[1] when obtaining a significant OR test may either conclude: (a) some instruments are invalid, (b) the model specification is deficient (so its implied errors may correlate even with instruments which would be valid for an adequate model specification), or (c) both model and instruments are unsound. We have shown that this shilly-shally is a consequence of not building on a firm maintained hypothesis. Under $\mathcal{M}_2^{L>K}$ an insignificant OR statistic is either due to validity of the tested additional moment conditions, or to lack of power of the test under (a), but not to (b) or (c), whereas a significant OR statistic is either due to a Type I error or to invalid instruments. Prac-

[1] See, for instance, Hayashi (2000, p.218) and Cameron and Trivedi (2005, p.277).

titioners who interpret an insignificant OR statistic as approval of the chosen model specification and the employed instruments should always realize that this presupposes validity of the untested hypothesis that this instrument set contains a subset which already just-identifies the coefficients of this chosen model specification.

The presented results on instrument validity and coefficient restriction tests can rather straight-forwardly be generalized to linear models with nonspherical disturbances estimated by generalized method of moments, because GMM conforms to IV/2SLS after proper transformations. When $\varepsilon \sim (0, \sigma_\varepsilon^2 \Omega)$ this transformation requires (as in GLS) to premultiply the model by an $n \times n$ matrix Ψ, where $\Psi'\Psi = \Omega^{-1}$, so that $\varepsilon^* = \Psi\varepsilon \sim (0, \sigma_\varepsilon^2 I)$, but employ now as instruments $Z^\dagger = (\Psi')^{-1}Z$, see Kiviet and Feng (2016). Hence, all above results can easily be reformulated for the linear GMM context, where (incremental) OR tests are usually addressed as Sargan-Hansen tests and indicated as (differences) in J tests.[2] In case of unknown heteroskedasticity (hence Ω is diagonal but not available), rather than robustifying inefficient IV, its findings $\hat{\beta}$ and $\hat{\varepsilon} = y - X\hat{\beta}$ can be employed in a second step to obtain asymptotically efficient feasible GMM inference. This can be represented[3] as follows. Let $y_i^* = y_i/\hat{\varepsilon}_i$, $X^* = (x_1^*, .., x_n^*)'$ with $x_i^* = x_i/\hat{\varepsilon}_i$ and $Z^\dagger = (z_1^\dagger, ..., z_n^\dagger)$ with $z_i^\dagger = z_i \times \hat{\varepsilon}_i$. Next regressing y^* on X^* using instruments $Z^\dagger$ yields $\hat{\beta}^\dagger = (X^{*\prime}P_{Z^\dagger}X^*)^{-1}X^{*\prime}P_{Z^\dagger}y^* = [X'Z(Z^{\dagger\prime}Z^\dagger)^{-1}Z'X]^{-1}X'Z(Z^{\dagger\prime}Z^\dagger)^{-1}Z'y$, with $Z^{\dagger\prime}Z^\dagger = \Sigma_{i=1}^n \hat{\varepsilon}_i^2 z_i z_i'$. Its variance could be estimated consistently by $\widehat{Var}(\hat{\beta}^\dagger) = [X'Z(Z^{\dagger\prime}Z^\dagger)^{-1}Z'X]^{-1}$, and writing $\hat{\varepsilon}^\dagger = y^* - X^*\hat{\beta}^\dagger$ the Sargan-Hansen statistic is $S^\dagger = \hat{\varepsilon}^{\dagger\prime}P_{Z^\dagger}\hat{\varepsilon}^\dagger = (y - X\hat{\beta}^\dagger)'Z(Z^{\dagger\prime}Z^\dagger)^{-1}Z'(y - X\hat{\beta}^\dagger)$, which is asymptotically equivalent to the GMM maximized criterion function. Hence, also for feasible GMM all earlier findings on CR and IOR tests have their (asymptotically equivalent) counterparts. Note, however, that a robustified variance estimator for $\hat{\beta}$ would be $(\hat{X}'\hat{X})^{-1}\Sigma_{i=1}^n \hat{\varepsilon}_i^2 \hat{x}_i \hat{x}_i'(\hat{X}'\hat{X})^{-1}$ with $P_Z X = \hat{X} = (\hat{x}_1, ..., \hat{x}_n)'$. Employing this in a robustified version of the CR test W_{β_2} of (2.4) does no longer seem to lead to a statistic which can be expressed equivalently as the difference between maximized robustified criterion functions, as in (2.9). Note that a robustified Sargan statistic for IV is given by $\hat{\varepsilon}'(\Sigma_{i=1}^n \hat{\varepsilon}_i^2 z_i z_i')^{-1}\hat{\varepsilon} = (y - X\hat{\beta})'Z(Z^{\dagger\prime}Z^\dagger)^{-1}Z'(y - X\hat{\beta})$. Due to the consistency of $\hat{\beta}$ this is asymptotically equivalent[4] to the 2-step statistic $S^\dagger$, and thus a robustified CF test has correspondences with a CR test based on 2-step feasible GMM estimator $\hat{\beta}^\dagger$ and no longer with one based on the inefficient $\hat{\beta}$.

Also in the extended contexts indicated in the preceding paragraph the practical relevance of the here presented findings is the following. An (incremental) Sargan-Hansen

[2] When embedded into a properly extended framework generalizations for nonlinear models can be obtained as well; see, for instance, Newey (1985).

[3] See Newey and West (1987, p.786).

[4] Baum et al (2003, p.18) claim that S^* and $S^\dagger$ are even numerically equivalent, which does not seem right.

test on the validity of (particular) orthogonality conditions uses exactly the same (or at least an under the null asymptotically equivalent) test statistic with the same asymptotic null distribution as a test for zero restrictions on the coefficients of additional regressor variables which may stem from a particular fairly wide group. However, in essence these two test procedures build on different nonnested maintained hypotheses which enables them to produce discriminative inferences.

References

Baum, C., Schaffer, M., Stillman, S., 2003. Instrumental variables and GMM: estimation and testing. *Stata Journal* 3, 1-31.

Cameron, A.C., Trivedi, P.K., 2005. *Microeconometrics: Methods and Applications.* Cambridge University Press.

Guggenberger, P., 2012. A note on the (in)consistency of the test of overidentifying restrictions and the concepts of true and pseudo-true parameters. *Economics Letters* 117, 901-904.

Hayashi, F., 2000. *Econometrics*. Princeton, Princeton University Press.

Hansen, P.L., 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029-1054.

Hendry, D.F., 2011. On adding over-identifying instrumental variables to simultaneous equations. *Economics Letters* 111, 68-70.

Kiviet, J.F., Feng, Q., 2016. Efficiency gains by modifying GMM estimation in linear models under heteroskedasticity. UvA-Econometrics working paper 2014/06.

Newey, W., 1985. Generalized method of moments specification testing. *Journal of Econometrics* 29, 229-256.

Newey, W.K., West, K.D., 1987. Hypothesis testing with efficient method of moments estimation. *International Economic Review* 28, 777-787.

Parente, P.M.D.C, Santos Silva, J.M.C., 2012. A cautionary note on tests of overidentifying restrictions. *Economics Letters* 115, 314-317.

Sargan, J.D., 1958. Estimation of economic relationships using instrumental variables. *Econometrica* 26, 393-514.

When is it really justifiable to ignore explanatory variable endogeneity in a regression model?

Jan F. Kiviet*

Version of 10 May 2016
JEL Classifications: *C2; C13; O5*

Keywords:
sensitivity analysis; simultaneity; asymptotic expansions; least-squares; growth regression

Abstract

A procedure that aims to pinpoint the sensitivity of ordinary least-squares based inferences regarding the degree of endogeneity of some regressors has been put forward in Ashley and Parmeter, *Economics Letters*, 137 (2015) 70-74. Here it is demonstrated that this procedure is based on an incorrect and systematically too optimistic asymptotic approximation to the variance of inconsistent least-squares. Therefore, and because the suggested sensitivity findings pertain to a random set of estimated endogeneity correlations, the claimed significance levels are misleading. For a very basic one coefficient model it is demonstrated why much more sophisticated asymptotic expansions under a stricter set of assumptions are required. This enables to replace some of the flawed sensitivity analysis results for an empirical growth model by findings based on a proper limiting distribution for a feasible inconsistency corrected least-squares estimator, given an adopted fixed range for the correlation between the endogenous regressor and the disturbance. Finally it is discussed how similar results could be achieved for models with several possibly endogenous regressors estimated by possibly endogenous instruments.

1. Introduction

An attempt is made in Ashley and Parmeter (2015a), henceforth APLS, to assess the sensitivity of ordinary least-squares (OLS) based tests with respect to simultaneity of an arbitrary number of regressors. The methods used in APLS are basically a specialization

*Emeritus Professor of Econometrics, Amsterdam School of Economics, University of Amsterdam, PO Box 15867, 1001 NJ Amsterdam, The Netherlands (j.f.kiviet@uva.nl). Most of this paper was written while I was enjoying hospitality as Visting Professor at the Division of Economics, School of Humanities and Social Sciences, Nanyang Technological University, Singapore. Comments by a referee are gratefully acknowledged.

of an approach put forward in Ashley and Parmeter (2015b), henceforth APIV, which aims to assess the sensitivity of instrumental variables (IV) based inference in situations where instruments are in fact endogenous. The APLS results are obtained by simply adopting in the APIV approach the regressors as instruments. The APIV study builds on Ashley (2009).

In Kiviet (2013), henceforth KLS, we developed an asymptotically valid inference method on the basis of inconsistent OLS results for models with one endogenous regressor. This would not be possible without making a further identifying assumption. In the KLS approach identification is achieved by making a nonzero moment condition, instead of the habitual zero moment conditions exploited by consistent IV. This approach enables to assess the sensitivity of OLS based t-test inferences over intervals nested in $(-1, +1)$ which should contain the true value of the correlation between the single endogenous regressor and the disturbance term.

APLS uses a similar, yet different, conversion of OLS inference than put forward in KLS. However, both approaches depart from an assessment of the limiting distribution of an unfeasible estimator which corrects OLS for its inconsistency. In this paper we will just focus on the results in APLS for which KLS offers an alternative, i.e. for models with just one endogenous regressor. We will show that the APLS approach suffers from two fundamental flaws. First, its asymptotic analysis is much too naive, which leads to three shortcomings: (i) it does not highlight that numerical assumptions on third and fourth moments are required; (ii) it overlooks that a choice has to be made regarding conditioning or not on exogenous regressors; (iii) the asymptotic variance adopted is systematically too small. Second, the suggested approach leads to an assessment of sensitivity over an adopted set of values for the covariance between regressors and disturbances. To interpret this for the dataset under study, the correlation between regressors and disturbances is estimated on the basis of the adopted covariance and the sample variances of the regressors and the residuals. However, the randomness of these latter estimates is not taken into account, whereas it is in the KLS approach. Therefore, test results obtained by the KLS approach are either tight or conservative, so they can claim asymptotic significance at a level equal to or not exceeding some chosen value, whereas the APLS results cannot.

By analyzing in all detail a very simple case we will show in Section 2 that obtaining the required asymptotic results is much more involved than suggested in APLS. In Section 3 we will compare for an empirical growth model with just one endogenous regressor the sensitivity results obtained in APLS by those produced by the KLS approach. In the concluding Section 4 we put the foregoing into perspective, and indicate some further literature that seems useful for achieving the goals of assessing the sensitivity of OLS and IV with respect to invalid orthogonality conditions for models with possibly more

than just one endogenous regressor.

2. Examination of a simple case

Consider the very simple linear regression model with just one random regressor and i.i.d. observations

$$y = x\beta + \varepsilon, \text{ with } x \sim (0, \sigma_x^2 I) \text{ and } \varepsilon \sim (0, \sigma_\varepsilon^2 I), \tag{2.1}$$

where both y and x are $n \times 1$ observed data vectors. Regressor x could be endogenous. Therefore we suppose that

$$x = \xi + \lambda\varepsilon, \tag{2.2}$$

where ε and $\xi \sim (0, \sigma_\xi^2 I)$ are independent. For $i = 1, ..., n$ we denote $\sigma_{x\varepsilon} = E(x_i\varepsilon_i) = \rho\sigma_x\sigma_\varepsilon = \lambda\sigma_\varepsilon^2$, thus endogeneity correlation $\rho = \lambda\sigma_\varepsilon/\sigma_x$ with $|\rho| < 1$. Of course, $\sigma_x^2 = \sigma_\xi^2 + \lambda^2\sigma_\varepsilon^2$, thus $\sigma_\xi^2 = (1-\rho^2)\sigma_x^2$. For $\lambda = 0$ (which implies $\rho = 0$) regressor x is exogenous and endogenous otherwise.

We make standard regularity assumptions, involving that the unknown scalars β, λ, $\sigma_x > 0$ and $\sigma_\varepsilon > 0$ are all finite. So $x'x > 0$ and the ordinary least-squares (OLS) estimator for β exists and is given by

$$b = x'y/x'x = \beta + x'\varepsilon/x'x. \tag{2.3}$$

We shall examine its limiting behavior for $n \to \infty$.

2.1. Expanding the OLS estimator

For $w_{x\varepsilon} = x'\varepsilon/n - \sigma_{x\varepsilon}$ we find $E(w_{x\varepsilon}) = 0$ and $Var(w_{x\varepsilon}) = n^{-2}\Sigma_{i=1}^n Var(x_i\varepsilon_i - \sigma_{x\varepsilon}) = n^{-2}\Sigma_{i=1}^n E[\xi_i\varepsilon_i + \sigma_{x\varepsilon}\sigma_\varepsilon^{-2}(\varepsilon_i^2 - \sigma_\varepsilon^2)]^2$. Assuming further that ε_i has a symmetric distribution with finite kurtosis coefficient κ (hence $E\varepsilon_i^3 = 0$ and $E\varepsilon_i^4 = \kappa\sigma_\varepsilon^4$) then $Var(w_{x\varepsilon}) = n^{-1}[\sigma_\xi^2\sigma_\varepsilon^2 + (\kappa-1)\sigma_{x\varepsilon}^2] = O(n^{-1})$. Thus, the order of probability of $w_{x\varepsilon}$ is $O_p(n^{-1/2})$, as $n^{1/2}w_{x\varepsilon}$ has a finite distribution. Likewise, assuming $E(x_i^3) = 0$ and $E(x_i^4) = \kappa\sigma_x^4$, we find for $w_{xx} = x'x/n - \sigma_x^2$ that it is $O_p(n^{-1/2})$. In fact,

$$w_{x\varepsilon} \sim (0, n^{-1}[1 + (\kappa-2)\rho^2]\sigma_\varepsilon^2\sigma_x^2) \text{ and } w_{xx} \sim (0, n^{-1}(\kappa-1)\sigma_x^4). \tag{2.4}$$

For the inverse of $x'x/n = \sigma_x^2(1 + \sigma_x^{-2}w_{xx})$ we find

$$\begin{aligned}(x'x/n)^{-1} &= \sigma_x^{-2}(1 + \sigma_x^{-2}w_{xx})^{-1} \\ &= \sigma_x^{-2}(1 - \sigma_x^{-2}w_{xx} + \sigma_x^{-4}w_{xx}^2 - \sigma_x^{-6}w_{xx}^3 + ...) \\ &= \sigma_x^{-2} - \sigma_x^{-4}w_{xx} + O_p(n^{-1}),\end{aligned} \tag{2.5}$$

where we took it for granted that all the omitted terms, which are of decreasing order, have a sum of the same order as the first and largest omitted term, which is $\sigma_x^{-6} w_{xx}^2 = O_p(n^{-1})$.

Substituting (2.5) we find for the OLS estimator (2.3)

$$\begin{aligned} n^{1/2}(b-\beta) &= n^{1/2}(x'x/n)^{-1}x'\varepsilon/n \\ &= (\sigma_x^{-2} - \sigma_x^{-4} w_{xx}) n^{1/2}(\sigma_{x\varepsilon} + w_{x\varepsilon}) + O_p(n^{-1/2}) \\ &= n^{1/2}\sigma_{x\varepsilon}/\sigma_x^2 + n^{1/2}(\sigma_x^{-2} w_{x\varepsilon} - \sigma_{x\varepsilon}\sigma_x^{-4} w_{xx}) + O_p(n^{-1/2}), \end{aligned}$$

where the term $-n^{1/2}\sigma_x^{-4} w_{xx} w_{x\varepsilon}$ could be absorbed in the remainder term because $w_{xx} w_{x\varepsilon} = O_p(n^{-1})$. This yields the expansion

$$n^{1/2}(b - \sigma_{x\varepsilon}/\sigma_x^2 - \beta) = n^{1/2}(\sigma_x^{-2} w_{x\varepsilon} - \sigma_{x\varepsilon}\sigma_x^{-4} w_{xx}) + O_p(n^{-1/2}), \tag{2.6}$$

for the infeasible consistent estimator $b - \sigma_{x\varepsilon}/\sigma_x^2$.

2.2. Limiting distributions

Because $\sigma_x^{-2} w_{x\varepsilon} - \sigma_{x\varepsilon}\sigma_x^{-4} w_{xx}$ can be written as the sample average of n i.i.d. random drawings with zero mean and finite variance a standard Central Limit Theorem can be applied. Therefore the limiting distribution of (2.6) will be normal. Its variance is given by the variance of the leading term of (2.6), which has a finite distribution. It can be found from (2.4) and from

$$\begin{aligned} E(w_{x\varepsilon} w_{xx}) &= E[(x'\varepsilon/n - \sigma_{x\varepsilon})(x'x/n - \sigma_x^2)] \\ &= n^{-2}E(x'\varepsilon x'x) - n^{-1}\sigma_{x\varepsilon}E(x'x) - n^{-1}\sigma_x^2 E(x'\varepsilon) + \sigma_{x\varepsilon}\sigma_x^2 \\ &= n^{-1}[2\rho + (\kappa - 3)\rho^3]\sigma_x^3\sigma_\varepsilon, \end{aligned} \tag{2.7}$$

as $E(x'\varepsilon) = n\sigma_{x\varepsilon}$, $E(x'x) = n\sigma_x^2$ and

$$\begin{aligned} E(x'\varepsilon x'x) &= E[\Sigma_{i=1}^n(\xi_i\varepsilon_i + \lambda\varepsilon_i^2)\Sigma_{j=1}^n(\xi_j^2 + 2\lambda\xi_j\varepsilon_j + \lambda^2\varepsilon_j^2)] \\ &= 2\lambda n\sigma_\xi^2\sigma_\varepsilon^2 + \lambda n^2\sigma_\varepsilon^2\sigma_\xi^2 + \lambda^3[n(n-1) + \kappa n]\sigma_\varepsilon^4 \\ &= \rho(1-\rho^2)(2n + n^2)\sigma_x^3\sigma_\varepsilon + \rho^3[n^2 + (\kappa - 1)n]\sigma_x^3\sigma_\varepsilon \\ &= n[(n+2)\rho + (\kappa - 3)\rho^3]\sigma_x^3\sigma_\varepsilon. \end{aligned}$$

This yields $Var(\sigma_x^{-2} w_{x\varepsilon} - \sigma_{x\varepsilon}\sigma_x^{-4} w_{xx}) = n^{-1}[1 + (2\kappa - 5)\rho^2 - (\kappa - 3)\rho^4]\sigma_\varepsilon^2\sigma_x^{-2}$, and thus, when $\kappa = 3$ (so especially when x_i and ε_i are jointly normal), then

$$n^{1/2}(b - \sigma_{x\varepsilon}/\sigma_x^2 - \beta) \xrightarrow{d} \mathcal{N}(0, (1-\rho^2)\sigma_\varepsilon^2/\sigma_x^2). \tag{2.8}$$

It seems that APLS in their formula (7), when specialized to the one regressor case, use the same infeasible inconsistency expression $\sigma_{x\varepsilon}/\sigma_x^2$ for the correction of the OLS

estimator.[1] However, this is not the case.[2] They consider in their formula (6) an estimator that in the present simple one regressor context is equivalent to

$$\tilde{b} = (x'y - n\sigma_{x\varepsilon})/x'x = \beta + w_{x\varepsilon}/(x'x/n). \tag{2.9}$$

Employing (2.5) we find the expansion

$$\begin{aligned} n^{1/2}(\tilde{b} - \beta) &= n^{1/2}(\sigma_x^{-2} - \sigma_x^{-4}w_{xx})w_{x\varepsilon} + O_p(n^{-1/2}) \\ &= n^{1/2}\sigma_x^{-2}w_{x\varepsilon} + O_p(n^{-1/2}). \end{aligned} \tag{2.10}$$

Since $var(n^{1/2}\sigma_x^{-2}w_{x\varepsilon}) = [1 + (\kappa - 2)\rho^2]\sigma_\varepsilon^2/\sigma_x^2$, for $\kappa = 3$ this yields limiting distribution

$$n^{1/2}(\tilde{b} - \beta) \xrightarrow{d} \mathcal{N}(0, (1 + \rho^2)\sigma_\varepsilon^2/\sigma_x^2). \tag{2.11}$$

Hence, the by APLS naively chosen expression $\sigma_\varepsilon^2/\sigma_x^2$ for the limiting variance of $n^{1/2}(\tilde{b} - \beta)$ is wrong, and under simultaneity it is systematically too small when $\kappa > 2$. Note that choosing (2.8) as a starting point seems much more attractive because its limiting distribution has a systematically smaller variance.

Due to skipping a formal derivation of (2.11) APLS do not realize either that assumptions on third and fourth moments are required. Moreover, they do not recognize that it matters whether or not one conditions on the random exogenous component ξ of the vector x. In KLS it has been shown that when conditioning on ξ, while assuming normality of x and ε, one obtains

$$n^{1/2}(b - \sigma_{x\varepsilon}/\sigma_x^2 - \beta) \xrightarrow{d} \mathcal{N}(0, (1 - \rho^2)(1 - 2\rho^2 + 2\rho^4)\sigma_\varepsilon^2/\sigma_x^2), \tag{2.12}$$

which is even more attractive than (2.8), because $1 - 2\rho^2 + 2\rho^4 \leq 1$ for $|\rho| < 1$. However, in what follows we will focus on the unconditional case; note that this will yield results which are conservative (over-cautious) for the conditional case.

For the unconditional case it has been derived in KLS as well that for the feasible consistent estimator b_ρ^*, which is obtained by correcting b using the actual value of ρ and consistently estimating σ_ε^2 and σ_x^2, one has

$$n^{1/2}(b_\rho^* - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma_\varepsilon^2/\sigma_x^2), \tag{2.13}$$

where

$$\begin{aligned} b_\rho^* &= b - n^{1/2}\rho(1 - \rho^2)^{-1/2}se(b), \text{ with} \\ se(b) &= s/(x'x)^{1/2} \text{ and } s^2 = (y - xb)'(y - xb)/(n - k). \end{aligned} \tag{2.14}$$

[1] Note that when $\Sigma_{X\varepsilon}$ is introduced in (6) of APLS there is a confusing typo: $E(X_i'\varepsilon_i)$ should read $E(X'\varepsilon)$.

[2] The term $-E_{XX}^{-1}\Sigma_{X\varepsilon}$ should be removed from formula (7) in APLS, because estimator $\tilde{\gamma}$ has already been corrected. However, this has been corrected by $-(X'X/n)^{-1}\Sigma_{X\varepsilon}$.

Here $se(b)$ is the expression for the usual OLS standard error estimate of b as produced under assumed exogeneity of x. In the present special model k (the number of regressors) is one, but for the asymptotic result to hold a degrees of freedom correction is irrelevant, of course. Estimator b^*_ρ is consistent for β, because $\sigma_{x\varepsilon}/\sigma^2_x = \rho(\sigma^2_\varepsilon/\sigma^2_x)^{1/2}$, whereas $x'x/n$ is consistent for σ^2_x and, as has been proved in KLS, $s^2/(1-\rho^2)$ is a consistent estimator of σ^2_ε. Note that $n^{1/2}se(b) = O_p(1)$ and therefore the inconsistency correction term in b^*_ρ is finite too and vanishes only for $\rho = 0$.

One may find it inappropriate that above estimator b^*_ρ is addressed as a feasible estimator, whereas in practice the actual value of ρ is commonly unknown. A not unreasonable response to that is: OLS/IV estimators are usually not labelled as unfeasible either, whereas their underlying (just) identifying orthogonality conditions simply adopt the value zero for the correlation between regressor(s)/instrument(s) and disturbance, for which it is equally difficult or sheer impossible to find statistical evidence, see Kiviet (2016).

Result (2.13) is quite remarkable, not because consistent feasible estimator b^*_ρ is found to have, due to estimating $(\sigma^2_\varepsilon/\sigma^2_x)^{1/2}$, a larger asymptotic variance than the unfeasible estimator $b - \sigma_{x\varepsilon}/\sigma^2_x$, but because this increment exactly leads to the same asymptotic variance as b has when it is consistent. However, this equivalence only holds (unconditionally) under the special assumptions made regarding skewness and kurtosis for x and ε; it can be shown that the asymptotic variance of b^*_ρ will be larger in case of excess kurtosis. It is remarkable too that estimating $(\sigma^2_\varepsilon/\sigma^2_x)^{1/2}$ for obtaining consistent estimator b^*_ρ increases the limiting variance of $b - \sigma_{x\varepsilon}/\sigma^2_x$ less than estimating just σ^2_x fo obtaining consistent estimator $\tilde{b}$.

Summarizing: Unlike KLS, APLS do not fix ρ but $\sigma_{x\varepsilon}$, which seems much less practicable, because this is not dimensionless like ρ is. Moreover, it requires to employ the less attractive limiting distribution (2.11) of consistent estimator $\tilde{b}$. Using simply, as APLS do, the incorrect standard expression $\mathcal{N}(0, \sigma^2_\varepsilon/\sigma^2_x)$, which does apply to b^*_ρ, will lead to systematically too optimistic inferences.

2.3. Sensitivity of OLS with respect to endogeneity

Asymptotically valid inference on β based on (2.13) can be obtained from the asymptotic approximation

$$b^*_\rho \overset{a}{\sim} \mathcal{N}(\beta, s^2/[(1-\rho^2)x'x]), \tag{2.15}$$

which for $\rho = 0$ simplifies to the standard result $b \overset{a}{\sim} \mathcal{N}(\beta, s^2/x'x)$. Let us focus now on testing $\mathcal{H}_0 : \beta = \beta_0$ against $\mathcal{H}_1 : \beta \neq \beta_0$ for known numerical value β_0. The usual statistic $t(\beta_0) = (b - \beta_0)/se(b)$ is asymptotically standard normal under $\mathcal{H}_0$, provided

$\rho = 0$. From (2.15) it follows that under simultaneity

$$t_\rho(\beta_0) = (1-\rho^2)^{1/2}(b^*_\rho - \beta_0)/se(b) = (1-\rho^2)^{1/2}t(\beta_0) - n^{1/2}\rho \tag{2.16}$$

is asymptotically standard normal under $\mathcal{H}_0$. This result can be used for a sensitivity analysis as follows. Imagine we find a t-value of 3 in a regression where $n = 100$. Then by solving $3(1-\rho^2)^{1/2} - 10\rho > 1.96$, which gives $\rho \leq 0.102$, we learn that rejection against a right-hand side alternative at the asymptotic significance level of 2.5% requires a negative or just a very small positive simultaneity correlation.

Note that for finite $t(\beta_0)$ and $n > 4$ we find that for ρ getting closer to $+1$ we have $t_\rho(\beta_0) < -c$ and for ρ getting closer to -1 we have $t_\rho(\beta_0) > c$, for c some positive critical value. This means that for any OLS test result and any chosen nominal significance level we can always find a range of values for the degree of simultaneity for which the test is asymptotically significant and also one for which it is insignificant. This clearly demonstrates the inadequacy of the APLS approach, because they report in their Table 1 that they did establish OLS inferences which were found to be robust to any degree of simultaneity ($r_{\min} = 1$).

As we understand it, the essential elements of the APLS sensitivity analysis in the context of the above simple one coefficient model involves the following. Suppose one has found for a particular data set that $|t(\beta_0)| > z_{1-\alpha/2}$, where z_p denotes for $0 < p < 1$ the p^{th} quantile of the standard normal distribution. So $\mathcal{H}_0$ is rejected at level α, presupposing exogeneity. Now the method seeks to establish the values ρ for which $\mathcal{H}_0$ would still be rejected. This is done as follows. The set of values for $\sigma_{x\varepsilon}$ (here scalar), say $\mathcal{S}(\sigma_{x\varepsilon})$, is assessed (by a random search procedure), giving $\mathcal{S}(\sigma_{x\varepsilon}) = \left\{\sigma_{x\varepsilon} \in \mathbb{R} : z_{\alpha/2} < (\tilde{b} - \beta_0)/se(\tilde{b}) < z_{1-\alpha/2}\right\}$, where

$$\begin{aligned} \tilde{b} &= b - \sigma_{x\varepsilon}/(x'x/n) \text{ with} \\ se(\tilde{b}) &= \hat{\sigma}_{\sigma_{x\varepsilon}}/(x'x)^{1/2} \text{ and } \hat{\sigma}^2_{\sigma_{x\varepsilon}} = (y - x\tilde{b})'(y - x\tilde{b})/n. \end{aligned} \tag{2.17}$$

When the numerical problem to assess $\mathcal{S}(\sigma_{x\varepsilon})$ has been solved, the corresponding set of values for $\hat{\rho}$ given by $\mathcal{S}(\hat{\rho}) = \left\{\hat{\rho} = \sigma_{x\varepsilon}/[\hat{\sigma}^2_{\sigma_{x\varepsilon}}(x'x/n)]^{1/2} : \sigma_{x\varepsilon} \in \mathcal{S}(\sigma_{x\varepsilon})\right\}$ is assessed, and the conclusion is drawn that rejection of $\mathcal{H}_0$ is robust with respect to endogeneity for all values covered by $\mathcal{S}(\hat{\rho})$.

Hence, a crucial difference with KLS is that a choice is made regarding $\sigma_{x\varepsilon}$ and not with respect to ρ directly. The price for that is that APLS find a random set $\mathcal{S}(\hat{\rho})$ for ρ, and omit to discuss the consequences of this randomness (which is not due to the random search, but to the dependence of $\hat{\rho}$ on $\hat{\sigma}^2_{\sigma_{x\varepsilon}}$ and on $x'x$). Moreover, as shown in the foregoing section, the formula used for $se(\tilde{b})$ in (2.17) is systematically too small.

3. Empirical illustration

The methods developed in APLS have been applied to a particular empirical growth model presented in Mankiw et al. (1992) where possible endogeneity of regressors has not been taken into account. This model for 98 countries can be represented as $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta x_{i4} + \varepsilon_i$, where y_i is per capita output, x_{i2} is the rate of human capital, x_{i3} is investment in physical capital and x_{i4} is the logarithm of the sum of the population growth rate, the growth rate in technology and the depreciation rate. The focus in the APLS analysis is on testing $\beta_2 = 0$ and $\beta_2 + \beta_3 + \beta_4 = 1$ (constant returns to scale). APLS find OLS results as given in their equation (8), which are slightly different from those in Mankiw et al. (1992, p.420, Table II, first column). APLS perform a sensitivity analysis for four different scenarios, which allow either x_{i2}, or x_{i3} or x_{i4} to be endogenous, or x_{i2} and x_{i3} are jointly assumed endogenous.

The KLS results for regression models with just one explanatory variable that may be endogenous can also serve the situation where the model has some further exogenous regressors, which have all been partialled out. This means that only for the first scenario of APLS (x_{i2} endogenous) we can produce KLS results relevant for testing β_2. For these access to the observations on the regressand and regressors is not required. The OLS coefficient estimates and standard errors presented in APLS equation (8) suffice to analyze $t_\rho(\beta_0)$ of (2.16). Just for illustrative purposes we also present in Table 1 results relevant on testing β_3 allowing x_{i3} to be endogenous and on β_4 allowing x_{i4} to be endogenous.

Table 1: KLS sensitivity of significance at 2.5% of growth coefficients

alternative hypothesis	$\beta_2 > 0$	$\beta_3 > 0$	$\beta_4 < 0$
admissible values for ρ	$\rho \leq 0.572$	$\rho \leq 0.311$	$\rho \geq -0.222$

The parallel finding in APLS regarding β_2 is: $r_{\min} = 0.425$ at 5% (from their text it is not clear whether they confront the test statistic with critical value 1.96 or 1.64; when we test the significance of β_2 at 5% critical value 1.64 the result is $\rho \leq 0.591$). Note that there is also a range of ρ values for which the observed OLS t-ratio's imply significance of opposite sign. For $\rho \geq 0.782$, which would lead to a huge positive bias of b and thus of $t(\beta_0)$, the empirical results corroborate with $\beta_2 < 0$ at 2.5%.

4. How to tackle more general cases?

From the above it should be clear that the KLS approach has yet been developed for only very few special simple cases, whereas the APLS approach has serious flaws in various of its underpinnings. Even when employing sound asymptotics, the latter will lead

to findings which are hard to interpret because the resulting $r_{\min}$ vector is random and obtaining an asymptotic approximation to its distribution to assess its accuracy seems far from easy. On the other hand, further development of the KLS approach for more general cases seems possible, but requires an asymptotic analysis which certainly cannot be characterized as in APLS (just above their section 2.2) as "easy" and "straightforward", because of the following three reasons.

First, although the corrected estimator (6) in APLS is clearly consistent, it is not self-evident what its limiting distribution will look like unless one embarks on a proper derivation. Doing so reveals that extra assumptions are required regarding the numerical values of third and fourth moments of the regressors and disturbance. Second, an inconsistent estimator has a bias which is $O(1)$, and hence any useful random assessment of this bias will be $O_p(1)$, and thus employing such an assessment to correct the inconsistent estimator will affect its limiting distribution and require a further expansion.[3] Third, a further complicating issue is that whereas the limiting distribution of standard consistent estimators is similar whether or not one conditions on exogenous variables, this situation apparently changes when consistency is achieved by correcting an inconsistent estimator.

Many aspects of these complications for regression models with an arbitrary number of endogenous and exogenous regressors have already been addressed in Kiviet and Niemczyk (2012) with respect to OLS estimation and in Kiviet and Niemczyk (2014) with respect to IV estimation when invalid instruments may have been used. A next step should be to obtain for these more general settings the limiting distributions of inconsistency corrected estimators which are feasible in the sense of KLS, and next exploit these to produce inference which is valid over a credible set of values for all endogeneity correlations. Only after this has been achieved for empirically relevant models it seems justifiable to provide a fully satisfactory answer to the question posed in the titles of APLS and of this study. Other approaches to achieve a similar goal in one way or another can be found in Murray (2006), Small (2007), Ebbes et al. (2009), Conley et al. (2012) and Kraay (2012).

References

Ashley, R., 2009. Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis. *Journal of Applied Econometrics* 24, 325-337.

[3] When a consistent estimator is biased, its bias is usually $O(n^{-1})$. The bias of the estimator can be reduced by subtracting an $O_p(n^{-1})$ assessment of this bias. In that case the corrected estimator retains the same limiting distribution as the uncorrected estimator, because the correction just affects higher-order asymptotic aspects.

Ashley, R.A., Parmeter, C.F., 2015a. When is it justifiable to ignore explanatory variable endogeneity in a regression model? *Economics Letters* 137, 70-74.

Ashley, R.A., Parmeter, C.F., 2015b. Sensitivity analysis for inference in 2SLS estimation with possibly-flawed instruments. *Empirical Economics* 49, 1153-1171.

Conley, T., Hansen, C, Rossi, P., 2012. Plausibly exogenous. *The Review of Economics and Statistics* 94, 260-272.

Ebbes, P., Wedel, M., Bockenholt, U., 2009. Frugal IV alternatives to identify the parameter for an endogenous regressor. *Journal of Applied Econometrics* 24, 446-468.

Kiviet, J.F., 2013. Identification and inference in a simultaneous equation under alternative information sets and sampling schemes. *The Econometrics Journal* 16, S24-S59.

Kiviet, J.F., 2016. Discriminating between (in)valid external instruments and (in)valid exclusion restrictions. UvA-Econometrics working paper 2015/04, revised 5 May 2016. Forthcoming in the *Journal of Econometric Methods*.

Kiviet, J.F., Niemczyk, J., 2012. The asymptotic and finite sample (un)conditional distributions of OLS and simple IV in simultaneous equations. *Journal of Computational Statistics and Data Analysis* 56, 3567-3586.

Kiviet, J.F., Niemczyk, J., 2014. On the limiting and empirical distribution of IV estimators when some of the instruments are actually endogenous, pp.425-490 in *Advances in Econometrics, Volume 33; Essays in Honor of Peter C.B. Phillips* (Eds. Yoosoon Chang, Thomas B. Fomby, Joon Y. Park), Emerald Group Publishing Limited, Bingley, UK.

Kraay, A., 2012. Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach. *Journal of Applied Econometrics* 27, 108-128.

Mankiw, N.G., Romer, D., Weil, D.N., 1992. A contribution to the empirics of economic growth. *Quarterly Journal of Economics* 107, 407-437.

Murray, M. P., 2006. Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives* 20, 111-132.

Small, D. S., 2007. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association* 102, 1049-1058.

A critical appraisal of studies analyzing co-movement of international stock markets with a focus on East-Asian indices

Jan F. Kiviet[*] and **Zhenxi Chen**[†]

Version of 5 February 2016
JEL Classifications: *C18, C52, C58, G15*
Keywords:
diagnostic tests, interlinkage of stock markets, maintained hypotheses, model specification, test methodology

Abstract

Literature is reviewed on the analysis of co-movement between the price indices of stocks or their realized revenues at various markets. Four major categories of frequently recurring methodological shortcomings are registered. These are: (i) omitted regressor problems, (ii) neglecting to verify agreement of estimation outcomes with adopted model assumptions, (iii) employing particular statistical tests in inappropriate situations and, occasionally, (iv) lack of identification. We focus on studies regarding the interlinkages between emerging and more developed stock markets, in particular those in East-Asia and Wall Street. The devastating effects of the detected methodological defects are explained analytically and also illustrated by simulations and empirical examples.

1. Introduction

Its economic and further social and technological developments over the last half century have put Asia more in the forefront in many respects. After the impressive economic achievements by Japan from early on in the second half of the 20th century, also the so-called 'Asian Tigers' (Hong Kong, Singapore, South-Korea and Taiwan) soon realized many decades of impressive growth figures, later followed in particular by China and

[*]Emeritus Professor of Econometrics, Amsterdam School of Economics, University of Amsterdam, P.O. Box 15867, 1001 NJ Amsterdam, The Netherlands (j.f.kiviet@uva.nl). This paper was written while I enjoyed hospitality as Visiting Professor at the Division of Economics, School of Humanities and Social Sciences, Nanyang Technological University, 14 Nanyang Drive, Singapore 637332.

[†]Faculty of Economics, Business and Social Sciences, Christian-Albrechts University, Olshausenstrasse 40, 24118 Kiel, Germany (z.chen@economics.uni-kiel.de)

India. Despite a few setbacks for some countries for a few specific periods, the unprecedented developments realized over the last few decades, and also the expectations they are calling with respect to the next era, made some opinion leaders already speak about the 'Asian Century'; see Mahbubani (2006) for further background.

The economic liberalization and acceleration have induced rapid developments at the South-East Asian capital markets and initiated their gradual integration within the global financial market. As a consequence, over the last few decades, many researchers examined the developments in stock markets for various Asian countries and their interlinkages with markets in their neighboring countries as well as with the major international financial centres. In these studies widely divergent research methods have been used to analyze and characterize these relationships and, not-surprisingly, often have arrived at conflicting conclusions. In this paper we will not focus in the first place on the different empirical findings reported by all these earlier studies, but on the various methodologies that they have used. After all, only inferences obtained from sound research methods should deserve serious further attention.

In the vast literature on the analysis of relationships between financial markets many studies focus on bivariate relationships, just involving two markets; others include more than two in a multivariate approach. Irrespective of the number of markets taken into consideration, the literature can be divided roughly into three different approaches regarding the chosen market characteristic of major interest. This is either: (i) the price index p_{it} of stock market i at time period t or its natural logarithm $\log(p_{it})$; or (ii) the conditional expectation $E(r_{it} \mid \mathcal{I}_t)$, where $r_{it} = \log(p_{it}) - \log(p_{i,t-1})$ is revenue and $\mathcal{I}_t$ represents an information set usually containing many (mostly lagged stock market) variables; or (iii) an aspect of the distribution of revenue other than its conditional mean, such as for instance the conditional centered second moment $Var(r_{it} \mid \mathcal{I}_t)$, or possibly particular quantiles of the density of r_{it}. An additional issue is usually whether the stock market prices should be taken in terms of local currency or be expressed in, for instance, US dollars. Many studies explore both variants and often find little difference between these two options.

For a recent historical survey on approaches to model individual stock returns see Koundouri et al. (2016). Studies that aim to explain movements in the actual price variables of different markets often employ cointegration analysis after nonstationarity of the stock indices has been assessed. Studies that chose to focus on modeling revenue (sometimes because no cointegration between indices could be established) usually specify some type of model for its conditional expectation and variance in terms of a dynamic relationship involving the revenues at various markets. Studies under (ii) often use a vector autoregressive (VAR) model and assume conditional homoskedasticity, whereas the studies under (iii) tend to use a very simple model for the conditional expectation and then employ many parameters for modeling the conditional heteroskedasticity. The latter approach may be induced by strong belief in the efficient market hypothesis, which suggests that the conditional expectation of revenue is constant. However, the approach under (ii) as a rule provides empirical evidence of its significant nonconstancy.

The majority of studies analyze daily closing values of the stock market indices. That different markets are located in different time zones is usually not taken into account among South-East Asian countries, because the trading hours have a substantial overlap. For obvious reasons, however, often the index of the preceding day is taken for the New York stock index when modeling its effects on Asian markets. However, as we shall clarify below, this can have a peculiar effect on Granger-causality tests. Weekend days are excluded from the data set and usually the index of the preceding day is used for the missing data on bank holidays, which are not always the same in all countries. Fewer studies on stock market interlinkage analyze weekly or monthly data.

From this literature a seriously disturbing observation has to be made regarding the progression of science. A major characteristic of this literature is that its contributors usually generously cite empirical findings obtained in earlier studies, irrespective of whether these have been obtained by a conflicting model specification or methodology. However, any discussion is then usually just about differences in the outcomes, instead of focussing on evaluating the pro's and con's of the chosen model specifications and the employed statistical techniques. In the review below, we will primarily highlight –and criticize where we think this is justified– the chosen methodology of a great (but by far not exhaustive) number of studies published in peer reviewed journals over the last two decades. In our opinion, the resulting empirical findings merit serious substantive consideration only, if the underlying methodology has no obvious flaws. Few studies seem to belong to the latter category.

In Section 2 we review many earlier studies on linkages between emerging and more developed stock markets, grouped in a series of subsections according to the three approaches distinguished above, and comment on their chosen methodology. There we note four major categories of problems which undermine many published empirical findings. All of them as a rule lead to seriously affected inferences, such as biased estimates of reaction coefficients and glaring misrepresentation of the actual significance level of test outcomes and of their reported p-values. These four categories of methodological errors are: (a) omitted regressor problems or not controlling adequately for relevant covariates; (b) neglecting to verify harmony of obtained estimation outcomes and the adopted model assumptions on which inference has been built; (c) employing particular statistical test procedures in inappropriate situations because the maintained hypotheses (a notion to be clarified below) do not hold; and, occasionally, (d) identification problems which preclude any useful statistical analysis. Further explanations of these rather technical issues and some relevant derivations are presented in their most simple formal format in a series of appendices. In Section 3 we provide some illustrations, both empirical and synthetic (by simulation), to highlight the devastating consequences of committing any of the four mentioned categories of methodological sins. Finally, some conclusions are drawn in Section 4.

2. Review of earlier results

Here we discuss a great number of published studies with relevance for the relationships between stock market indices or revenues. We will categorize all these studies with respect to their bivariate or multivariate nature, and regarding the methodology used for analyzing the chosen characterization of the dependent variable. In particular we will verify how careful all these studies are regarding four rather fundamental aspects for any empirical statistical analysis. These four (not fully disjunct) aspects are: (a) has it formally been examined –especially in case of a bivariate analysis– whether the adopted approach may suffer from omitted variables bias, because an appropriate specification may require a more extended multivariate approach, see also Appendix A; (b) does the study use diagnostics in order to verify whether or not the statistical assumptions made are actually supported by the empirical findings, see also Appendix B; (c) are the statistical tests used suitable for the particular situation in which they are applied (see Appendix C for an example of an often misused test), and (d) are the estimated parameters formally identified (see Appendix D).

We group the cited articles in separate subsections which collect studies that focus on the level of indices, on modeling the expectation of revenue, and on other aspects of the distribution of revenue respectively. In each subsection these reviews are presented in chronological order of the examined publications.

2.1. Studies with focus on the level of stock indices

Chung and Liu (1994) analyze weekly stock market data expressed in local currencies from early 1985 until May 1992 for the US, Japan, Taiwan, Hong Kong, Singapore and South Korea. Nowhere in their analysis the Black Monday stock market crash of 1987 plays a role. They do not allow for any (possibly temporary) structural changes in any of their models. They find all series to be nonstationary (after taking their natural logarithm). Next they apply the Johansen methodology to establish the number of stochastic trends and cointegration relationships. The numbers they find prove to be highly dependent on the chosen lag length in the VEC (vector error-correction) model. Their maximum likelihood based inferences require normality of the disturbances. However, normality is rejected for all individual countries. Though, by using alternative multivariate normality tests (which involve a much larger number of degrees of freedom) the meshes of the net are stretched so far that establishing misspecification escapes. Using high degree of freedom Ljung-Box tests the authors declare all their estimated equations "free of serial correlation problems", which disregards the poor power of such tests. Next a criterion based on the average absolute forecast errors over the last year (taken out of the estimation sample) is employed to select the model which uses 12 lags, instead of 24 or 36. Finally, this model (which imposes without testing constant coefficients and constant error variances over the whole sample) is further used to characterize and to test various hypotheses on the long and short run dynamics of the six analyzed stock market indices.

Arshanapalli et al. (1995) use more classic (Engle-Granger) cointegration and error-correction analysis to examine the stock market linkages between East-Asian markets (excluding China) and the US. Based on daily data ranging from January 1986 to May 1992, and analyzing separately the data pre and post October 1987, they conclude that the cointegrating structure of these markets and their dependence on the US increased after October 1987. However, their multivariate error-correction regressions actually result from reformulated and restricted VAR models with maximum lag order 4 for the 7 considered stock market indices, in which one linear cointegration relationship has directly been imposed. This cointegration relationship has been obtained from a static constant parameter regression between all indices, in full trust that its putative super-consistency will repel any bias due to neglected simultaneity and omitted short-run dynamic adjustments. This imposition undermines the interpretability of the presented inflated t-statistics. Apparently not noticing that no alternatives were allowed in their analysis, the authors nevertheless infer that "the six major Asian stock exchanges are linked together with a long-run equilibrium relationship". First, however, by diagnostic testing (see Appendix B) the adequacy of the dynamic specification of the error-correction regressions should have been verified, especially by tests for omitted regressors (see Appendix A). Also LM (Lagrange multiplier) tests for (higher-order) serial correlation should have been used, and not just the DW statistic, which requires all regressors to be strictly exogenous and hence is inappropriate in error-correction models. Presentation of results for the static long-run regressions over shorter subperiods might have given genuine support to the claimed existence of different though constant parameter long-run relationships before and after October 1987.

Ghosh et al. (1999) present a classic (Engle-Granger) cointegration analysis too. They restrict themselves to separate bivariate models for nine South-East Asian developing stock market indices. In all these bivariate models the other variable is the developed stock market of either the US or Japan. They use 141 daily data for estimation and 60 for out of sample prediction. Estimation is just based on the end of March until end of October 1997 period. None of the model formulations used does in any way explicitly examine whether the parametrization should pay respect in this way or another to the fact that in July 1997 the Asia financial crisis broke out in Thailand. Hence, all models impose many untested restrictions. Moreover, in the methodology practiced in this paper rejection of stationarity of the errors of a static bivariate cointegration relationship is automatically interpreted as absence of cointegration, whereas it is quite possible that cointegration would be established after including other more or less developed markets as well. The obtained error correction model estimates are all interpreted as if they represent the true data generating process, although no diagnostic test outcomes that would support this have been presented. The authors even claim that their findings imply the existence of causality; apparently they wrongly take this to be equivalent to correlation. Finally, they show that out of sample predictions by their models are superior to naive (no change) forecasts, which does not surprise. They would better have compared their predictions with the actually observed realizations in such a

way, that by a Chow-type predictive failure test a genuine diagnostic on the quality of their findings had been obtained.

Huang et al. (2000) use daily price data from 1992 to 1997 for the US and various Asian stock markets. Allowing for structural breaks, they reject (except for Shanghai and Shenzen) bivariate cointegration between the data for log of price. Next they examine Granger causality by bivariate dynamic regressions in the revenue data (though, surprisingly, not allowing for any structural breaks here). From bivariate VAR(k) models, where the Schwartz criterion is used to select lag order $k = 1$, they claim to find unidirectional Granger causality from the US to Hong Kong and no Granger causality in either direction between Hong Kong and Shanghai, nor between the US and Shanghai. The limitations of bivariate Granger tests, where in case of omitted regressors the included regressors serve as proxies for the omitted regressors leading to estimation bias (see Appendix A), are not mentioned. Nor has any evidence been presented on the statistical adequacy of the dynamic regressions in the form of diagnostics on serial correlation, heteroskedasticity, structural breaks or omitted (lagged) regressors.

Cheng and Glascock (2005) analyze weekly stock market price indices denominated in US dollars for China, Hong Kong, Taiwan, Japan and the US, covering January 1993 to August 2004. First, they demonstrate for the three markets from the Greater China Economic Area (GCEA) that each is not weakly efficient, because predictions by simple random walk models are dominated by those obtained from univariate models using more information from the past. Next, various single equation Johansen-type cointegration tests follow, but excluding the case where both Japan and US could appear with all the GCEA states in one cointegration relationship. Cointegration is not found, neither when tests are conducted over two (not specified) consecutive subperiods. Of course, it would have made sense here to split the two subperiods such that they deliberately exclude the weeks strongly associated with the Asian financial crisis of 1997/1998 in order to avoid contagion effects. Next, models are estimated by linear least-squares which are simple transformations of first-order bivariate autoregressive distributed lag equations. Overlooking the dangers of omitted relevant other markets or of longer lags, and not realizing that validity of the earlier inferences would imply nonstationarity of the error terms, no single diagnostic is presented. Nevertheless, from standard t-tests far fetched conclusions are being drawn regarding existing nonlinear pairwise relationships. Finally, an innovation accounting analysis follows, now based on estimating a multivariate VAR in the revenues. It has not been reported how lag lengths were assessed, nor whether diagnostics produced any evidence supporting the chosen specification.

Yang et al. (2006) study the interdependence among the stock markets of US, Germany and four major Eastern European emerging countries. Based on daily data from early 1995 to mid 2002 they first apply constant parameter cointegration analysis to the price series measured in logs of the market indices. This is done separately to three pre-crisis and three post-crisis years, although later, by using a two-year rolling window, strong evidence is presented that even in noncrisis years both the long-run and the short-run reactions vary over time. Nevertheless, the authors claim that the estab-

lished cointegration relationships, where the lag length in the VAR systems have been selected on the basis of the Akaike information criterion, do pass LM tests for first- and fourth-order autocorrelation. The empirical conclusions are mainly based on the so-called persistence profile technique and on generalized forecast error variance decompositions. These numerical measures have been obtained from models which impose doubtful parameter constancy, and are next interpreted just on the basis of their point estimates, without taking their (obviously hard to assess) standard errors into account.

Huyghebaert and Wang (2010) are well aware that different research methodologies may be at the base of the mixed results in many earlier studies. For seven East Asian stock markets and the US they use daily data from mid 1992 to mid 2003, distinguishing the mid-1998 until mid-1999 crisis year from the pre and post crisis periods, examining two variants for the latter. They choose for a multivariate VAR framework for the log of stock-exchange indices when performing cointegration tests, assuming constant parameters over the four distinguished periods. They expect –overenthusiatically– "to correctly specify" their various VAR models by using a likelihood-ratio test for lag length to "ensure that all dynamics in the data are being captured". Unlike Yang et al. (2006) for Eastern-Europe, they find for East-Asia cointegration only during the crisis period, but not before, nor some time after. Next, Granger-causality tests are performed based on a VEC specification in the log indices for the crisis period and on VAR specifications in the revenues for the other periods. Apparently the authors do not realize that cointegration directly implies causality. In their causality analysis parameter constancy over the subperiods is taken for granted, no results for serial correlation tests are being mentioned. As it seems, only F-test outcomes and their p-values for joint significance of all lag coefficients for the separate countries are being mentioned, neglecting that just one significant single coefficient would imply Granger-causality too. Moreover, in line with the (incorrect) interpretation of insignificant outcomes of Sims' likelihood ratio test as indicating truth of the null hypothesis, F-test outcomes with p-values just above 5% are now interpreted as establishing absence of Granger-causality. Whether or not the presented tests are robust for (untested) heteroskedasticity is not clear. This is all followed by a meticulous generalized impulse response analysis, which of course is built on –and is thus conditional on the not exhaustively tested validity of– the earlier established specifications. The short-term causal relationships inferred from them in terms of responses to one unit shocks are of course all random estimates, but presented without their standard errors these can barely be interpreted.

Burdekin and Siklos (2012) use daily data from early 1995 to mid 2010 to investigate the interaction between the Chinese, US and Asia-Pacific equity markets. Advocating an eclectic approach, they in fact produce a wide range of results which are mutually contradictory and all untenable from a statistical point of view. Seemingly meant to just present some initial simple stylized facts of the revenues, such as estimates of pairwise unconditional correlations and of first-order autoregression coefficients, next significance tests of these are presented too. However, proper interpretation of the tests on correlations (see Appendix C) require assumptions which have to be rejected given the different

coefficients obtained in the autoregressions. But also the latter are built on assumptions, such as no higher-order dependence and parameter constancy, which have not been verified. Next contagion tests are being produced in a set of fully symmetric equations [their (1.1)] which omit any dynamics but explicitly include endogenous regressors without imposing any restrictions which guarantee identification of the parameters of this simultaneous system. Although it is not mentioned which technique has been used to estimate the coefficients and their standard errors, due to the shortcomings just mentioned it should be obvious that no sensible interpretation of these results is possible either (see Appendix D). This is followed by estimating a dynamic conditional correlation (DCC) model. However, how the conditional expectation of the returns has been modeled this time, and whether this appeared satisfactory, is not mentioned, so probably again has not been verified. All attention goes here to the conditional covariance of the returns without any concern how this might be affected when the conditional expectation has not been modeled adequately. Finally static quantile cointegration regressions are estimated and their coefficients are interpreted on the basis of standard t-ratio's, which is inappropriate again.

Glick and Hutchison (2013) investigate China's financial linkages with Asia in both bond and equity markets using daily data from mid 2005 to end of 2012. After some standard unit root analysis just differenced data are being considered, starting off with wrongly (see Appendix C) employing significance tests on simple correlation coefficients of revenues. Next, a series of static bivariate, multivariate and pooled regressions are presented which expose extreme naivety regarding the properties of least-squares estimators and requirements to usefully interpret significance tests. Assuming for the moment that the simple multivariate regressions in their Table 2 are not misspecified, each significant coefficient rejects validity of a restricting bivariate regression, in which the presented significance test results are thus inappropriate, even if the regressors are hardly correlated because they are both serially dependent, given the causality tests presented in their Table A2. Moreover, the multivariate regressions also reject validity of the pooled regression, because the slope coefficients are clearly different for most of the individual Asian countries. However, the multivariate regressions are uninterpretable themselves too, because it seems unlikely that equity return in China is exogenous for all other Asian stock markets, so least-squares estimators are biased, not only due to possibly neglected reverse causality, but also due to omitted regressors because of neglected dynamics. Next, results are presented which undermine all findings presented before, because evidence is produced indicating that regression coefficients are nonconstant through time. However, that evidence is just derived from pooled regressions, which involve restrictions on the slopes which had already been shown to be incredible. Hence, the many alternative regressions presented in this study simply illustrate that none of them fully respects the actual interdependencies between the variables examined. None of the empirical conclusions drawn in this paper is based on solid statistical evidence.

Wang (2014) examines the interdependence and causality among six East-Asian stock

markets and the US. This is mainly a repetition of the Huyghebaert and Wang (2010) study, now using more recent daily data, namely from mid 2005 to mid 2013, thus focusing now on the effects of the 2008 global financial crisis instead of the 10 years earlier Asian financial crisis. Using the same methodology, the same criticism applies. In particular, without proper testing, parameter constancy is assumed within the three distinguished periods: pre-crisis, crisis and post-crisis. It seems that lag length tests have been performed rather mechanically, treating all East-Asian countries similarly, whereas possible misspecification of the VAR and VEC models has not been tested. The causality tests, for which it is not clear whether they are robust for heteroskedasticity, are joint on all lagged coefficients of each country, which obscures particular forms of causality.

2.2. Studies regarding stock revenue data

Some of the above mentioned studies, already turned to analyzing revenue after stock indices were found to be not cointegrated, at least over particular subperiods. Other studies turn to analyzing revenues right away. Within this group of studies we can distinguish those that focus primarily on modelling the conditional expectation of the distribution of revenue and those that focus primarily on other aspects of this distribution.

2.2.1. Focus on conditional expectation

Groenewold et al. (2004) analyze the dynamic interrelationships between the share markets of Shanghai, Shenzhen, Hong Kong and Taiwan. Although correctly criticizing similar earlier research using just bivariate cointegration, they overlook that their own approach may suffer from the same criticism as they exclude the international global financial market from their analysis. They use daily data from end of 1992 until end of 2001 and also consider two sub-samples, namely before and after the crisis period 1997/98. No cointegration can be established, except for the two markets in mainland China (Shanghai and Shenzhen) over the first sub-sample. Next they combine the two markets in mainland China and estimate trivariate VAR(5) models which pass tests for 1st and 4th order serial correlation. Knowing from various other studies that at least the Hong Kong and Taiwan markets are affected by the New York stock exchange all VAR inference may be seriously hit by omitted variable bias, and so will then be the resulting impulse responses and forecast error-variance decompositions, which lack a measure of precision (standard errors) anyway.

Zhu et al. (2004) analyze causal linkages between revenues at the stock markets of Shanghai, Shenzhen and Hong Kong, using daily data from early 1993 to the end of 2001. Not being able to find cointegration between the three indices, and not questioning whether this might be due to omissions such as the effect of global stock market indices and possible parameter nonconstancies, they move on to analyze causality between the three revenue series, with and without removing so-called calendar effects. They split the data period into two and estimate for each and the whole data period

VAR models. Optimal lag lengths have been determined mechanically by the Schwartz criterion. No diagnostic checks have been used, so coefficient restriction tests may suffer seriously from adopting invalid maintained hypotheses. The causality tests are only based on joint F-tests, using the 5% significance level as a razor blade, whereas p-values of individual t-tests could have given useful supplementary evidence. A similar analysis is also performed on the absolute values of the revenue series to analyze causality in volatilities.

Singh et al. (2010) analyze data for 15 stock markets, including US, Canada, UK, Germany, France and ten major Asian markets. They use daily revenue data over the 2000 through 2008 period obtained from indices both at opening and at closing of the market. First they estimate VAR(15) models using just one lag for both close-to-close returns and for open-to-open returns. No evidence on parameter constancy nor on lack of error serial correlation is provided. Next the two VAR models are adapted in the following way. For the two markets that open (close) earliest in the day (Japan and Korea) the VAR specification is unaltered (includes 15 lagged explanatories). The three markets that open (close) next are Singapore, Malaysia and Taiwan. For them in the VAR specification the lagged revenues of Japan and Korea are replaced by current values. And for the remaining ten VAR equations all lagged revenues at earlier opening (closing) markets are replaced by current revenues, and in addition the second through fifth lag of the dependent variable has been added, but not for France and Germany. The authors do realize that this introduces reverse causality (simultaneity) in their specifications (except for Japan and Korea), at least when the revenues of the earlier opening (closing) markets have been realized in part during overlapping opening hours, but they suggest (footnote 5) that this has only a minor effect. Instead of using an alternative and still consistent estimation methods they just use least-squares and wrongly interpret standard t-values in the standard way. For what reasons suddenly lag lengths larger than one have been included is not clarified. For none of these results the maintained hypotheses implicitly adopted given the estimation method used seem credible. Finally, chucking all earlier results for the conditional first moment of revenue, AR(1)-GARCH(1,1) models are presented, which do of course require a completely different set of maintained hypotheses, but these have not been tested either.

Jayasuriya (2011) examines interlinkages of stock returns for China and three of its emerging neighbors, namely Thailand, Indonesia and Philippines. Although citing various similar studies mentioning the significant role of the US, this study chooses to include only some control variables from the real economy, namely the growth rates in interest, inflation and exchange rates, which –without any further motivations– are all treated as exogenous. The analysis is based on the monthly stock returns from November 1993 to July 2008, to which Jarque-Bera tests have been employed, which is incorrect, because it is most unlikely that these returns are serially uncorrelated and have time-invariant first four moments, which is a requirement for this test to apply. Because the revenues are found to be stationary it is claimed that a VAR model is appropriate, overlooking that it should be examined as well whether a stationary error-

correction term in the stock indices has to be included as well. All conclusions are based on VAR(1) models which have passed LM serial correlation tests at lags 12 and 24. Of course, order 1 and 2 tests should have been used as well, because if the problems are especially at those lags, the tests employed lack power. Parameter constancy is taken for granted, one simple dummy is supposed to account for the Asian financial crisis, and the omission of more developed and other neighboring stock markets is not supposed to lead to estimation bias. Hence, the presented impulse response functions and variance decompositions are clearly based on much too weak grounds.

Chow et al. (2011) analyze weekly returns and their absolute value (which they take as their measure of volatility) from 1992 until 2010 at the stock markets of Shanghai and New York. In their bivariate analyses they exclude, without testing, any possible effects from Hong Kong or any other East Asian markets, so no evidence has been produced to refute obvious criticism that further explanatory variables have wrongly been omitted undermining all their inferences. First some simple descriptions of the (absolute) returns at the two markets are presented in the form of means, variances and simple correlations, over the two decades separately, and also over the full 20 years period. These suggest nonconstancy, but later analyses (and earlier results in Chow and Lawler, 2003) indicate that these three data characteristics are in fact nonconstant within the two decades too, so apparently they just should be seen as rough indicators for which no precision measure can be given. Next (in their Table 3) two simultaneous and unidentified (see Appendix D) equations have been estimated (also over the two decades separately), and wrongly it has been inferred from two t statistics (which are algebraically equivalent, see Appendix F) that the New York market affects Shanghai and vice versa. Next they estimate time-varying coefficient models (not mentioning that if these are found to be appropriate all their earlier inferences have to be withdrawn) using Kalman filter techniques and making strong distributional assumptions on error terms and weekly random movements of the coefficient values. Plausibility of these assumptions has not been verified from the fitted models by the authors. Results demonstrating some of the serious shortcomings of these bivariate time-varying coefficient models can be found in Chen et al. (2015).

2.2.2. Focus on (partial) correlations

Aityan et al. (2010) is a relatively recent study in a long tradition (see further references in their paper) which is purely based on calculating simple bivariate correlation coefficients of the revenues at two markets over subperiods. In this particular paper they use daily data and calculate correlations for separate years. The very serious shortcoming of this analysis is that interpretation of these correlations would only be possible if for both series the daily data are serially uncorrelated and have constant mean and variance. However, all studies discussed in the foregoing subsection demonstrate that these conditions are not satisfied. Hence, all statistical tests produced in this literature are invalid. The correlations presented do not establish estimates of a constant underlying population parameter. An indication of their accuracy by an appropriate standard error can not be given. Proper analysis of these revenue series demonstrates that more

sophisticated multivariate models with many more parameters are required in order to produce inferences that possibly make sense. Only for such more extended models their adopted maintained hypotheses are not easily refuted by empirical evidence.

Kenett et al. (2012) use partial correlations and thus extend the above criticized approach to a trivariate analysis. Also they use a much smaller window of a month and consider overlapping windows. However, our major criticism still applies. The approach does not start off from formulating a stochastic model for the interdependencies between the analyzed revenue series. Therefore, there is no maintained hypothesis for which tenability can be verified by testing it against alternatives. Also specific hypotheses regarding the obtained results cannot be tested statistically because the reliability of the estimates cannot be expressed in standard errors.

2.2.3. Focus on conditional variance and further aspects

Li (2007) presents a multivariate study based on daily revenue data from 2000 till mid 2005 for the S&P 500 and the stock markets of Shanghai, Shenzhen and Hong Kong. The mean of these four variables is simply modeled by a VAR(1) allowing its innovation errors to have a conditionally heteroskedastic covariance matrix. For the latter various forms have been tested, leading to the acceptance of an unrestricted four-variable asymmetric GARCH specification inspired by the BEKK parametrization. For $m = 4$ markets this has (next to four intercepts) $m^2 = 16$ coefficients regarding the one period lagged revenues for modeling the mean, and $3m^2 = 48$ parameters to model the error variance. Tenability of the assumption that the errors are innovations is only supported by high-order (12 and 24) Ljung-Box statistics. Their p-values (all above 0.15) inspire the author to conclude straightforwardly that the mean has been adequately specified. This conclusion seems rather rash, taking into account that usually no satisfactory parsimonious constant parameter specification can be found for the mean of revenues. The Ljung-Box test is well known for being heavily undersized (much too low rejection probability) and having poor power. A more serious test for the chosen specification of the mean should examine the significance of higher-order lagged revenues and possible structural changes in their values. Also the choice for daily data, given the diverging closing hours and disjunct bank holidays of the markets in the US and in China, may have curious effects on the obtained residuals. A poorly specified mean equation will lead to residuals which do not really represent true innovations and therefore may suggest findings regarding volatility linkage which will not be genuine but mere artefacts. Anyhow, on top of concluding to very mild unidirectional return linkages of mainland China with Hong Kong and of Hong Kong with the US, the study also claims to establish various rather subtile volatility linkages between the four markets. The latter are of course all conditional on the maintained hypothesis that the mean of the returns has been modeled adequately indeed by a constant parameter VAR(1). This supposition becomes more doubtful after the author has reduced the sample size in order to get rid of the 9/11 shocks. Re-estimation of the same model over the three years after 9/11 yields estimates for the mean of the revenues which still support unidirectional dependence of

Hong Kong on the US, but the weak dependence of mainland China on Hong Kong is no longer significant.

Zhang et al. (2009) analyze data for the Shanghai and Hong Kong composite stock indices over the period mid 1996 until late 2008. They start off with the bivariate equivalent of the specification used by Li (2007), oddly enough after they have already established the higher-order autoregressive nature of the revenue series for Hong Kong. Regarding return spillovers from Hong Kong to Shanghai and vice versa the same conclusions are drawn as in Li (2007), but not regarding volatility linkage, which they ascribe to the different data period. Although that may certainly be one of the reasons, it is strange that no remark has been made on the devastating effects of their choice for a bivariate analysis. This forces zero values for parameters that are significantly different from zero in Li's multivariate study. That Zhang et al. (2009) start off from a misspecified model for the mean of revenue is also communicated by the significant Ljung-Box statistics, which are presented in the tables, but receive no attention in the text at all. Nevertheless, Zhang et al. (2009) continue to use their limited specification of the mean model in a more sophisticated copula-based analysis. Although evidence is put forward that for Shanghai the linear autoregressive model for the mean improves by allowing for a nonlinear dynamic extension, again clear signs are presented (though again neglected) indicating that the resulting residuals of this extended mean equation do not represent a series of innovations, which implies that the copula approach has been built on an inappropriate likelihood function.

Beirne et al. (2010) construct GDP weighted weekly returns for the global market (US, Japan and four major European countries) and for large regions (including Asia) and next perform trivariate analyses of global and regional spillovers in emerging stock markets by specifying VAR-GARCH(1,1)-in-mean processes with BEKK representation. They impose, without testing, that there is no spillover from local markets to regional and global markets, nor from regional markets to the global market. The only diagnostic used is the ten lags Ljung-Box test, which rejects for only a few countries. The study misses results for the much more critical test on significance of the second and higher order lag in the VAR specifications. The results for the Asian countries are based on data from mid 1993 until early 2008. No checks on parameter constancy have been reported.

Li (2012) examines the stock market linkages between China, Korea, Japan and US. Using daily data for revenue from mid 1992 to early 2010 he applies a 4x4 GARCH-BEKK model and conditional correlation to uncover the linkages. The results suggest that China is linked to the oversea markets. Here too the GARCH-BEKK model has a specification of just one lag in the mean equation, which is accepted by the author given Ljung-Box test results at 12 and 24 lags. Hence, the same criticism applies as to Beirne et al. (2010). Given the results in the papers focussing on the expectation of revenue, just using a VAR(1), especially if it concerns daily data, does not seem right. In that case, the residuals do not represent innovations, which would turn all GARCH-BEKK findings into mere artefacts.

Zhang and Li (2014) use daily data from 2000 until 2012 on the Shanghai stock exchange composite index and the Dow Jones industrial average index. Allowing for a structural break they cannot establish cointegration between the indices; their explanations for that do not include that any bivariate approach seems inappropriate. Next they focus on analyzing revenues. First univariate models are used combining an ARMA(1,1) specification for the conditional expectation of revenue with GARCH(1,1) for the error terms, and next the two established standardized residual series are used in a trend analysis with structural breaks of the so-called conditional correlations, which they forget to define properly. Given the almost complete lack of diagnostic testing of the many specification assumptions made we suppose that the results should be interpreted much more cautiously than the authors do. Finally, in a quantile regression model it is suddenly supposed that the relationship between the two revenues requires in fact only 3 coefficients (between end of 2007 until early 2012), whereas the earlier analysis used at least 20 or more.

3. Illustration of consequences of poor econometrics

In this section we demonstrate the consequences of various methodological flaws by some simple simulations and also by empirical illustrations.

3.1. Some simulation findings

To demonstrate pitfalls associated with using the Jarque-Bera (JB) test for normality on revenue data and of using Ljung-Box (LB) or Box-Pierce (BP) tests for serial correlation in VAR models we simulated data from a very simple bivariate VAR(1) model where the intercepts in both equations are zero. The disturbance of the first equation $u_t^{(1)}$ may be ARMA(1,1) with parameters ρ and θ, and it also has a dummy explanatory variable d_t with coefficient β. The second equation is a simple AR(1) process. Further details on this system are that

$$y_t^{(1)} = 0.1y_{t-1}^{(1)} + 0.1y_{t-1}^{(2)} + \beta d_t + u_t^{(1)} \tag{3.1}$$

$$y_t^{(2)} = 0.2y_{t-1}^{(2)} + u_t^{(2)}, \tag{3.2}$$

for $t = 1, ..., n$, where

$$\begin{aligned} u_t^{(1)} &= \sigma_1[(1-\rho^2)/(1+2\rho\theta+\theta^2)]^{1/2}\xi_t \\ \xi_t &= \rho\xi_{t-1} + \varepsilon_t^{(1)} + \theta\varepsilon_{t-1}^{(1)} \\ u_t^{(2)} &= \sigma_2\varepsilon_t^{(2)}. \end{aligned}$$

We take $\sigma_1 = 0.01$ and $\sigma_2 = 0.05$, whereas the series $\varepsilon_t^{(1)}$ and $\varepsilon_t^{(2)}$ are mutually independent serially uncorrelated drawings from the standard normal distribution. Hence, disturbance $u_t^{(2)}$ has variance σ_2^2 and the AR(1) series $y_t^{(2)}$ has standard deviation $0.051 = 0.05/\sqrt{1-0.2^2}$. Since series ξ_t is an ARMA(1,1) process with variance $(1+2\rho\theta+\theta^2)/(1-$

ρ^2), disturbance $u_t^{(1)}$ is an ARMA(1,1) process with variance σ_1^2. This yields series $y_t^{(1)}$ and $y_t^{(2)}$ which reasonably mimic some of the basic properties of actual revenue series.

We did choose $y_{-50}^{(1)} = y_{-50}^{(2)} = 0$ and discarded the obtained warming-up values for $t = -50, ..., -1$ in estimation. We generated this system 10000 times for just a few specific values of β, ρ, θ and sample size n. Each replication we applied particular tests at significance level $\alpha = 0.05$ and counted the frequency of rejection. The JB test for normality, which is χ_2^2 distributed under its null hypothesis, we applied to the two synthetic revenue series. Tests against p^{th} order serial correlation for $p \in \{1, 10\}$ we applied to the first equation when estimated by least-squares assuming (sometimes incorrectly) that the disturbances $u_t^{(1)}$ are serially uncorrelated. Next to the BP and LB tests, which are χ_p^2 distributed under their null, we examined the LM and the LMF tests, which under the null are χ_p^2 and $F_{p,n-k-p}$ distributed[1], where k is the number of regressors in the equation under the null hypothesis.

All these tests are asymptotic tests, so for n finite their actual type I error probability may deviate from 0.05. If it is much larger than 0.05 for a particular test then a rejection could both be due to validity and invalidity of the null hypothesis, which renders the test not very useful. If it is much smaller than 0.05 the probability to commit a type I error is much lower than intended, which results in much larger probability of type II errors and thus in low power of the test. So, in both cases (over-rejection and under-rejection under the null) the ability of the test to help choosing between the null and the alternative hypothesis is negatively affected.

We first focus on the popular Jarque-Bera normality test. This has been developed for series which are (asymptotically) IID (independent and identically distributed) like regression residuals when the model is adequately specified and the errors are IID themselves. Hence, when employed to a (demeaned) revenue series, the JB test can only fruitfully be interpreted if this revenue series has first and second unconditional moments that are time-invariant, and these revenues are serially uncorrelated. However, these three characteristics are usually rejected empirically, implying that the maintained hypothesis of the JB test procedure does not hold. Hence, in such circumstances one should not employ the test. Nevertheless, we will use it to both variables $y^{(1)}$ and $y^{(2)}$ of the above system to examine its behavior. The dummy variable d_t has all its observations zero, except $d_{n/2} = 1$ (and we made sure n is even). Hence, the coefficient β models a one-off shock halfway the observed sample. We examined $\beta = 0$ and some negative values, equal to 2, 4 and 8 times σ_1, which represent a one day stock market crisis.

Table 1 presents the rejection probability by JB when applied to series $y^{(1)}$ in case $\rho = \theta = 0$. We note that for $-0.02 \leq \beta \leq 0$ values remarkably close to the chosen significance level are found, but for $\beta \leq -0.04$ the null hypothesis is much more often rejected, apparently simply because the mean is nonconstant. From these results one could wrongly conclude that the serial dependence of the series seems not a crucial issue. However, this is due to the moderate values of the coefficients of the lagged regressors

[1] We considered the versions which set pre-sample values of the least-squares residuals equal to zero.

in the system. We also applied the JB test to the $y_t^{(2)}$ series and found a rejection probability slightly below 0.05. However, upon increasing the AR(1) coefficient from 0.2 to 0.9 the rejection probability increased to 0.25 for the smaller and to 0.45 for the larger sample size. Hence, this highlights that the JB test should not be applied to revenue series if evidence is available that these are not IID, but only to the residuals of a well specified model for revenues.

Table 1: Rejection probabilities of JB normality test for series $y^{(1)}$

	$n = 150$				$n = 500$			
	β				β			
	0.00	-0.02	-0.04	-0.08	0.00	-0.02	-0.04	-0.08
JB	0.041	0.057	0.371	0.999	0.043	0.054	0.269	0.997

When analyzing the serial correlation tests we removed the dummy variable when generating and estimating the first equation. The rejection probabilities that we calculated can be found in Table 2. Note that under the null hypothesis (when $\rho = 0$ and $\theta = 0$) both the BP and LB test show profound size problems, which are opposite in nature for $p = 1$ and $p = 10$. Only for very large samples and p substantial too the control over type I errors is appropriate. For small p we note that size control gets even significantly worse for these two tests when n gets larger. Apparently the BP and LB rejection probabilities can be non-monotonic in n, which is a very bad sign. On the other hand, at the sample sizes examined both LM and LMF show no serious size problems. For all tests the rejection probability increases when the null is not true, as it should. From the results for the Lagrange-Multiplier tests we note that when the serial correlation problem is of low order (as in the examined AR(1) and MA(1) cases) power of the test benefits when using a low order alternative. The often much higher reject values of the BP and LB tests when the disturbances are AR(1) or MA(1) are meaningless, because the corresponding probabilities are also much higher than for LM and LMF when the null hypothesis is true.

Table 2: Rejection probabilities of residual serial correlation tests in (3.1)

	$n = 150$			$n = 500$		
	$\rho = 0.0$ $\theta = 0.0$	$\rho = 0.2$ $\theta = 0.0$	$\rho = 0.0$ $\theta = 0.2$	$\rho = 0.0$ $\theta = 0.0$	$\rho = 0.2$ $\theta = 0.0$	$\rho = 0.0$ $\theta = 0.2$
BP1	0.250	0.678	0.730	0.351	0.945	0.972
BP10	0.001	0.039	0.050	0.042	0.647	0.746
LB1	0.253	0.682	0.733	0.352	0.945	0.972
LB10	0.001	0.044	0.055	0.042	0.650	0.749
LM1	0.051	0.190	0.232	0.054	0.497	0.622
LM10	0.042	0.074	0.087	0.050	0.210	0.269
LMF1	0.049	0.184	0.226	0.054	0.495	0.619
LMF10	0.042	0.075	0.088	0.050	0.210	0.270

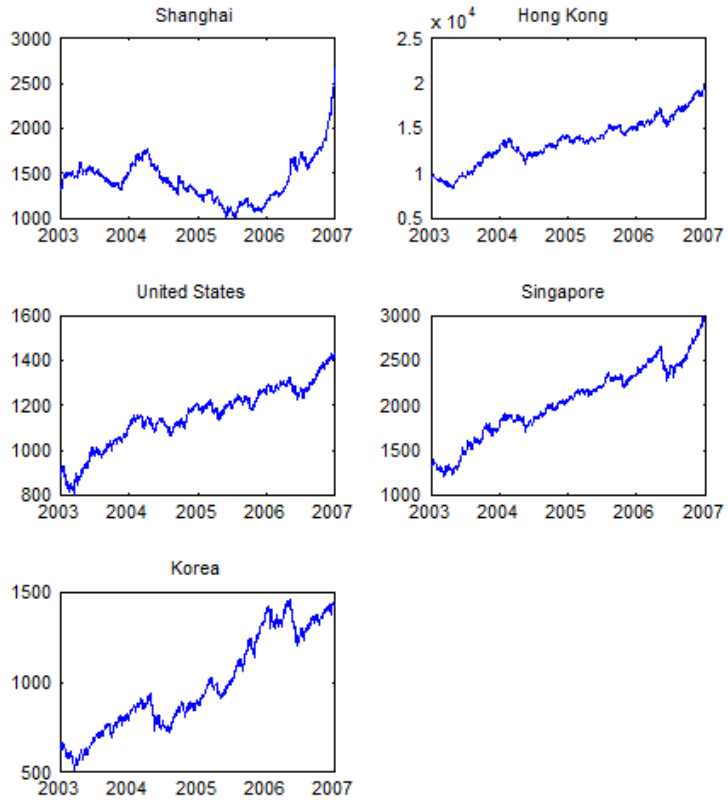
From these results one cannot predict what the patterns will be for similar or different values of n in larger VAR systems with longer lags and different coefficient values, but they do make clear that in the cases investigated there is little to choose between the BP and LB tests and that both can have large and small rejection probability under both the null and the alternative hypothesis, which completely undermines their signpost function. The LM and LMF tests show very reasonable size control, but their results also expose that one should certainly not overestimate the actual power of serial correlation tests. Hence, for none of the tests examined, it will ever make much sense to boldly conclude – as many practitioners wrongly do – that there is no serial correlation problem when the test employed happens not to reject.

3.2. Some empirical illustrations

In this subsection we use empirical data to illustrate the occurrence and serious consequences of various possible methodological flaws when estimating VAR models and performing tests for unit roots, cointegration and for causality. We perform these analyses by using the EViews package, and calculate the following diagnostics: (i) the BP (now referring to Breusch-Pagan) heteroskedasticity test, which tests whether the disturbance variance seems functionally related to the regressor variables; (ii) the JB (Jarque-Bera) normality of the residuals test; and (iii) also three variants of serial correlation tests, namely LB (Ljung-Box), LM (Lagrange multiplier χ^2 test), and LMF (Lagrange multiplier F-test). Each serial correlation test will be conducted at lag orders running from 1 up to 10.

The empirical data used are the daily stock indices of SH (Shanghai's SSE composite index), HK (Hong Kong's Hang Seng), US (S&P 500), SG (Singapore's STI) and KR (Korea's KOSPI index) over the period January 2003 till the end of 2006, all extracted from Yahoo finance. All these stock indices are expressed in local currencies. Deliberately the sample period (just over 1000 observations) is chosen such that financial crises periods are avoided, in order to keep away from extra hard model specification problems. To take into account the different time zones, for the US market the index of the preceding day has been taken. For all markets we use the preceding available observation for a bank holiday. Figure 1 plots the time series of the five indices. By visual check, we do not observe pronounced breaks except for Shanghai, for which there seems a structural break in the course of 2005.

Figure 1: Time-series of the five stock indices: Shanghai (SH), Hong Kong (HK), United States (US), Singapore (SG), Korea (KR)



3.2.1. Univariate unit root analysis

First, we will illustrate inference problems which may occur in the context of standard unit root analysis. We have estimated univariate autoregressive models for the natural logarithm of all the five stock indices at lag length $l = 0, 1, 2...$ with an intercept plus (or without) linear trend, and performed (augmented) Dickey-Fuller tests over the full sample and over the first and last part of the sample.

In Table 3 we report for illustrative purposes the results for Singapore and in Table 4 for the US. We present the test statistic (DF) and its corresponding 5% critical value (DFC), and also the p-value for each of the five diagnostics, where serial correlation is tested both for order 2 and order 10. In addition, the minimum p-value of the LMF test is reported and also the lag (≤ 10) at which it was obtained. Note that the unit root test presumes that the autoregressive model is well specified, meaning that the error terms should be serially uncorrelated, have time-invariant variance and are normally distributed. The top panel of the tables include the linear trend and the lower panel excludes it. The full sample results are for 2003 through 2006, whereas sub-sample "first" is for 2003-2004 and "last" for 2005-2006.

Table 3: Dicky-Fuller unit root tests and p-values of some diagnostic tests in AR(l) models for the natural logarithm of the Singapore stock index

intercept and linear deterministic trend included

sample						— order 2 —			— order 10 —			– Min –	
	l	DF	DFC	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
full	0	-2.40	-3.41	0.00	0.00	0.61	0.60	0.60	0.28	0.28	0.28	0.13	6
	1	-2.44	-3.41	0.00	0.00	0.81	0.57	0.57	0.32	0.25	0.26	0.13	6
	2	-2.41	-3.41	0.00	0.00	1.00	0.71	0.71	0.34	0.16	0.17	0.07	6
	3	-2.49	-3.41	0.00	0.00	1.00	0.86	0.86	0.29	0.17	0.18	0.06	4
	4	-2.50	-3.41	0.00	0.00	1.00	0.35	0.36	0.33	0.01	0.01	0.01	8
first	2	-2.08	-3.42	0.00	0.00	1.00	0.53	0.53	0.76	0.59	0.60	0.27	1
	3	-2.13	-3.42	0.00	0.00	1.00	0.47	0.48	0.80	0.60	0.61	0.23	4
last	2	-1.20	-3.42	0.00	0.00	0.99	0.08	0.08	0.07	0.06	0.06	0.03	5
	3	-1.18	-3.42	0.00	0.00	0.99	0.06	0.06	0.07	0.06	0.06	0.01	3

intercept included, no trend

sample						— order 2 —			— order 10 —			– Min –	
	l	DF	DFC	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
full	0	-0.64	-2.86	0.00	0.00	0.60	0.60	0.60	0.28	0.28	0.28	0.15	6
	1	-0.63	-2.86	0.00	0.00	0.72	0.50	0.50	0.30	0.25	0.25	0.13	7
	2	-0.67	-2.86	0.00	0.00	1.00	0.63	0.63	0.34	0.17	0.17	0.08	7
	3	-0.75	-2.86	0.00	0.00	1.00	0.76	0.76	0.30	0.17	0.17	0.08	7
	4	-0.69	-2.86	0.00	0.00	1.00	0.45	0.46	0.32	0.01	0.01	0.00	7
first	2	-0.93	-2.87	0.00	0.00	1.00	0.57	0.58	0.74	0.58	0.59	0.31	1
	3	-1.02	-2.87	0.00	0.00	1.00	0.46	0.46	0.76	0.60	0.61	0.24	1
last	2	0.49	-2.87	0.00	0.00	0.99	0.11	0.11	0.09	0.08	0.08	0.04	8
	3	0.51	-2.87	0.00	0.00	0.99	0.05	0.05	0.10	0.08	0.08	0.02	3

For Singapore, both the information criteria of Akaike and of Schwartz (AIC and SIC) prefer lag length zero, irrespective of the inclusion of a trend. At the same time, however, we note that the disturbances seem heteroskedastic and nonnormal (this is also found without first taking logs of the index). When a less parsimonious dynamic model is estimated the BP and JB tests do not improve, but some evidence of higher order serial correlation is detected as well. Neglecting the diagnostics, none of the DF values of Table 3 forces to reject the unit root hypothesis. However, the results also illustrate that the positive statement "the Singapore index is a unit root process" does not find firm support from these findings, also because of the substantial differences between DF values over the sub-samples.

From the results on the US presented in Table 4 we find again many rejections by the BP and JB tests. Irrespective of the inclusion of a linear trend the SIC criterion favors $l = 0$ whereas AIC prefers $l = 7$, whereas also the LMF tests reject low order autoregressive specifications. Again, for none of the rows in the table the unit root hypothesis is rejected, but at the same time no specification passes all the diagnostic

tests. This table illustrates as well that as a rule the Ljung-Box test is much too mild, and that the two information criteria are of little use for finding an adequate model specification.

Table 4: Dicky-Fuller unit root tests and p-values of some diagnostic tests in AR(l) models for the natural logarithm of the US stock index

intercept and linear deterministic trend included

sample						— order 2 —			— order 10 —			– Min –	
	l	DF	DFC	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
full	0	-2.82	-3.41	0.00	0.00	0.02	0.01	0.01	0.04	0.03	0.03	0.01	7
	1	-2.61	-3.41	0.00	0.00	0.34	0.17	0.17	0.23	0.02	0.02	0.00	7
	2	-2.42	-3.41	0.00	0.00	0.95	0.06	0.06	0.44	0.13	0.13	0.04	7
	3	-2.50	-3.41	0.00	0.00	0.95	0.02	0.02	0.46	0.05	0.05	0.01	7
	7	-2.27	-3.41	0.00	0.00	1.00	0.93	0.93	1.00	0.49	0.50	0.41	9
first	2	-1.99	-3.42	0.00	0.00	0.94	0.07	0.07	0.67	0.33	0.34	0.07	2
	3	-2.08	-3.42	0.00	0.00	0.95	0.05	0.06	0.71	0.22	0.23	0.05	3
last	2	-2.80	-3.42	0.00	0.08	1.00	0.75	0.75	0.95	0.58	0.59	0.41	8
	3	-2.75	-3.42	0.00	0.08	1.00	0.75	0.75	0.95	0.52	0.53	0.24	5

intercept included, no trend

sample						— order 2 —			— order 10 —			– Min –	
	l	DF	DFC	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
full	0	-1.12	-2.86	0.00	0.00	0.01	0.01	0.01	0.02	0.01	0.01	0.00	7
	1	-1.06	-2.86	0.00	0.00	0.25	0.14	0.15	0.16	0.01	0.01	0.00	7
	2	-0.85	-2.86	0.00	0.00	0.96	0.10	0.10	0.35	0.10	0.10	0.03	7
	3	-0.91	-2.86	0.00	0.00	0.96	0.01	0.01	0.36	0.04	0.04	0.01	7
	7	-0.81	-2.86	0.00	0.00	1.00	0.86	0.87	1.00	0.68	0.69	0.62	9
first	2	-0.59	-2.87	0.00	0.00	0.94	0.08	0.08	0.64	0.28	0.29	0.08	1
	3	-0.67	-2.87	0.00	0.00	0.94	0.04	0.04	0.66	0.19	0.20	0.04	1
last	2	-0.21	-2.87	0.00	0.04	1.00	0.80	0.80	0.85	0.33	0.33	0.19	8
	3	-0.16	-2.87	0.00	0.00	1.00	0.79	0.79	0.86	0.27	0.28	0.15	8

3.2.2. Cointegration analysis

Under the (somewhat doubtful) supposition that the natural logarithm of all five stock index series are genuine unit root processes, we next will illustrate inference problems which may easily occur in the context of cointegration analysis. Therefore we estimate VAR(l) models, with overall lag length $l = 0, 1, 2, ...$, first for an initial number of market indices ($m = 2$) and next on an extended number ($m > 2$). For all tables we will again report which lag length is preferred by either the AIC or SIC criterion. The tables present the p-values of both Johansen's Maximum Eigenvalue test ($\lambda_{\max}$) and the Likelihood Ratio trace test (LR_{tr}) to assess the number of cointegration relationships (CRs). As in Tables 3 and 4 we also report the p-value for each of the five diagnostics, where serial correlation tests are again given for order 2 and order 10, but also the

lag order for which the minimal p-value of LMF has been obtained. Now we use the multivariate versions of the diagnostics as provided by EViews. This means that the approximate critical values for the system LB test are taken to be χ^2 with $m^2(h-l)$ degrees of freedom, where m is the dimension of the VAR(l) system and h the order of serial correlation tested (here $h = 1, ..., 10$), so the test is only feasible for $h > l$. The LM type tests test $m^2 h$ restrictions. As Eviews does not provide the LMF test, we follow the procedure of Edgerton and Shukur (1999) to implement the LMF test. Note that the Johansen Maximum-Likelihood approach presumes that the VAR model is well specified, meaning that the error terms are serially uncorrelated, have time-invariant variance and are normally distributed. Like all studies reviewed in Section 2.1 we will allow for an intercept in the VAR(l) models but not for a linear deterministic trend. We will again report results for the full and for the two sub-samples.

Table 5: P-values of Johansen cointegration tests and diagnostics in VAR(l) models with intercept (without trend) for the natural logarithm of the stock indices of Hong Kong and Singapore.

						— order 2 —			— order 10 —			– Min –	
l	CRs	$\lambda_{\max}$	LR_{tr}	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
	full sample												
0	0	0.17	0.04	0.00	0.00	0.14	0.20	0.03	0.49	0.36	1.00	0.02	1
	≤ 1	0.07	0.07										
1	0	0.15	0.06	0.00	0.00	0.53	0.18	1.00	0.81	0.41	1.00	0.34	1
	≤ 1	0.13	0.13										
2	0	0.13	0.04	0.00	0.00	NA	0.92	1.00	0.87	0.36	1.00	1.00	1
	≤ 1	0.09	0.09										
8	0	0.36	0.12	0.00	0.00	NA	0.18	0.63	0.39	0.24	0.83	0.02	6
	≤ 1	0.09	0.09										
	first sub-sample												
8	0	0.84	0.73	0.00	0.00	NA	0.30	0.47	0.17	0.03	0.85	0.03	7
	≤ 1	0.48	0.48										
	last sub-sample												
8	0	0.16	0.09	0.00	0.00	NA	0.59	1.00	0.92	0.92	1.00	1.00	1
	≤ 1	0.20	0.20										

Table 5 presents results for bivariate VAR(l) models ($m = 2$) for the natural logarithm of the indices of Hong Kong and Singapore. BP and JB reject all examined VAR(l) models, suggesting heteroskedasticity and non-normality. AIC and SIC suggest the appropriate lag order to be 2 and 0 respectively for the full sample. The LR_{tr} test perceives one cointegration relationship over the whole sample at $l = 0$. However, this result should not be trusted due to the serial correlation. Evaluating at $l = 2$ as suggested by AIC, there is no longer serious serial correlation, and now the LR_{tr} test also indicates one cointegration relationship. However, at $l = 8$ no cointegration relationship is found in both full sample and sub-samples, and both full sample and first sub-sample suffer from serious serial correlation according to the LMF test.

Table 6 presents cointegration test results for trivariate VAR(l) models ($m = 3$) by adding Korea to the two markets system consisting already of Hong Kong and Singapore. BP and JB are again significant at all lag orders. Here both AIC and BIC favor the lag order to be 0, at which no cointegration relationship is detected. Concern for serial correlation is alleviated for the full sample as all p-values are above the 10% level, which is not the case for the first sub-sample at $l = 8$. Note that no cointegration is found, which of course undermines the results at lag order 2 in Table 5.

Table 6: P-values of Johansen cointegration tests and diagnostics in VAR(l) models with intercept (without trend) for the natural logarithm of the stock indices of Hong Kong, Singapore and Korea.

						— order 2 —			— order 10 —			– Min –	
l	CRs	$\lambda_{\max}$	LR_{tr}	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
	full sample												
0	0	0.54	0.27	0.00	0.00	0.48	0.45	0.73	0.70	0.72	1.00	0.12	1
	≤ 1	0.33	0.28										
	≤ 2	0.39	0.39										
1	0	0.46	0.32	0.00	0.00	0.91	0.42	1.00	0.86	0.68	1.00	1.00	1
	≤ 1	0.53	0.42										
	≤ 2	0.38	0.38										
2	0	0.39	0.22	0.00	0.00	NA	0.78	1.00	0.87	0.64	1.00	1.00	5
	≤ 1	0.41	0.31										
	≤ 2	0.36	0.36										
8	0	0.84	0.58	0.00	0.00	NA	0.28	0.46	0.54	0.57	1.00	0.42	1
	≤ 1	0.45	0.42										
	≤ 2	0.50	0.50										
	first sub-sample												
8	0	0.91	0.87	0.00	0.00	NA	0.12	0.42	0.09	0.19	1.00	0.00	1
	≤ 1	0.88	0.78										
	≤ 2	0.50	0.50										
	last sub-sample												
8	0	0.48	0.36	0.00	0.00	NA	0.42	1.00	0.92	0.97	1.00	0.95	1
	≤ 1	0.67	0.46										
	≤ 2	0.29	0.29										

In Table 7 we add Shanghai and the US to the system. As before BP and JB are significant at all lag orders. Now AIC and SIC suggest orders of 2 and 0 again. For these lag orders the LR_{tr} test indicates no and one cointegration relationship respectively. Note that at these two lag orders there are serious serial correlation problems according to the LM and LMF tests, thus these cointegration results should not be trusted. The serial correlation results also indicate clearly that the quality of inference regarding the appropriate lag order by the information criteria is most doubtful. Increasing the lag order, still no cointegration relationship is found, whereas at lag order 8 serial correlation problems seem no longer present. Investigation of the two sub-samples at lag order 8 indicates zero and two cointegration relationships, although the first sub-sample again

suffers from serious serial correlation. Hence, we still did not establish genuine long-run cointegration.

Table 7: P-values of Johansen cointegration tests and diagnostics in VAR(l) models with intercept (without trend) for the natural logarithm of the stock indices of Hong Kong, Singapore, Korea, Shanghai and the US

						— order 2 —			— order 10 —			– Min –	
l	CRs	$\lambda_{\max}$	LR_{tr}	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
	full sample												
0	0	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.47	0.00	1
	≤ 1	0.24	0.12										
	≤ 2	0.60	0.31										
	≤ 3	0.39	0.29										
	≤ 4	0.34	0.34										
2	0	0.21	0.07	0.00	0.00	NA	0.02	1.00	0.24	0.70	1.00	0.59	7
	≤ 1	0.23	0.23										
	≤ 2	0.75	0.53										
	≤ 3	0.57	0.46										
	≤ 4	0.40	0.40										
8	0	0.29	0.13	0.00	0.00	NA	0.27	1.00	0.73	0.75	1.00	0.59	1
	≤ 1	0.32	0.30										
	≤ 2	0.80	0.57										
	≤ 3	0.65	0.45										
	≤ 4	0.31	0.31										
	first sub-sample												
8	0	0.79	0.64	0.00	0.00	NA	0.00	0.59	0.05	0.46	1.00	0.00	1
	≤ 1	0.83	0.69										
	≤ 2	0.62	0.67										
	≤ 3	0.85	0.78										
	≤ 4	0.57	0.57										
	last sub-sample												
8	0	0.00	0.00	0.00	0.00	NA	0.78	1.00	0.72	0.54	1.00	1.00	1
	≤ 1	0.00	0.02										
	≤ 2	0.70	0.65										
	≤ 3	0.67	0.67										
	≤ 4	0.64	0.64										

At this stage we want to draw the following conclusions from the above. Since information criteria take it for granted that a VAR(l) model specification is adequate for the markets considered, they clearly should not be trusted to determine the lag order for performing cointegration tests when at the same time the specification does not pass diagnostic tests. However, no battery of diagnostics can ever guarantee that a particular specification sufficiently respects the actual data generating process. From the cointegration result of Table 7 using $l = 8$ and the last sub-sample one may conclude that if any cointegration relationship between these four Asian markets exists, this relationship

most probably must also involve the US. So note that this implies that all inferences in Tables 5 and 6, even those with satisfactory diagnostics, are misleading. Apparently, none of its test results can be associated with any degree of credible significance. This is evidenced in Table 8 which lists the two established normalized cointegration relationships in the last sub-sample for the five-markets model at $l = 8$. The US is significant for the two cointegration relationships. Hence, note that the results on the five markets destroy the trustworthiness of the results on fewer markets. Probably the same would happen with the five markets model if we added further markets.

Table 8: Normalized cointegration relationships as perceived from the five markets models (std. errors between parentheses)*

sample	Hong Kong	Singapore	Korea	Shanghai	US	Constant
last sub-sample	1	0	0.24	0.03	-4.60***	21.14***
			(0.20)	(0.14)	(0.90)	(4.61)
		1	0.28**	0.16*	-3.90***	16.92***
			(0.13)	(0.09)	(0.58)	(2.98)

* Significance at level 10, 5 and 1% is indicated by: *, ** and ***.

An issue of further concern for VAR modeling, not mentioned in any of the studies referred to above, is an uncomfortable consequence of the time zone difference between US and Asia. We can illustrate this for a simple binary VAR(1) system $y_t = Ay_{t-1} + \varepsilon_t$, where, for instance, $y_t = (s_{us,t-1},\ s_{sg,t})'$ contains the log stock indices of the US and Singapore. This entails the equations

$$\begin{cases} s_{us,t-1} = a_{11}s_{us,t-2} + a_{12}s_{sg,t-1} + \varepsilon_{1t}, \\ s_{sg,t} = a_{21}s_{us,t-2} + a_{22}s_{sg,t-1} + \varepsilon_{2t}. \end{cases}$$

Due to the habitual one period lag applied to the series for the US, the first equation is as intended, because the realization of $s_{sg,t-1}$ precedes the generation of $s_{us,t-1}$. However, due to taking this lag, the equation for Singapore becomes unsatisfactory, because for explaining $s_{sg,t}$ the latest available information on the US, being $s_{us,t-1}$, is not utilized. Consequences of this for VAR models in revenues we will examine in the next subsection.

3.2.3. Causality analysis

We continue to illustrate the devastating effects on inference when using underspecified models and now turn to Granger-causality tests. For that purpose we estimate VAR(l) models for $l = 0, 1, 2, ...$ on revenues (the first difference of the natural logarithm of the stock index) and again examine the effects of increasing the number of included markets $m = 2, 3, 5$. All equations have an intercept but not a linear trend (because its presence would be unacceptable according to finance theory). We just give results for the full sample size. For cases where existence of cointegration relationships is very likely the VAR model in revenues omits the "error correction term", so ideally it should not pass diagnostic checks and it should not be used for Granger-causality testing (also because joint dependence is already implied by cointegration). The tables report results for

each of the dependent variables (Dep) of the system in turn. For each group of lagged regressors the p-value of the F-test for their joint-significance is given, followed by the smallest p-value for the significance test of an individual lagged variable, followed by the lag order at which it occurs. Again results on the five diagnostics are given in the usual fashion, although, unlike what we did in the preceding subsection, we now present their single equation versions (so they do not involve residuals from the other equations). Due to the concern regarding the treatment of different time zones, we will (when the US is included in the VAR system) also present results for separate single equation regressions which always have as regressors the most recently already realized revenues at the other markets and their lags.

In Table 9 bivariate VAR results are given for Hong Kong and Singapore, for which Table 5 provided some dubious evidence of possible cointegration. SIC and AIC suggest the lag order of the VAR to be 0 and 1 respectively. Since some t-values are significant lag-order zero is clearly not acceptable. JB is always significant. However, for the F and t causality tests nonnormality of the disturbances should not be a serious problem, given the size of the sample. At $l = 1$, we detect no heteroskedasticity, and evidence of one causality relation is found, namely that Singapore Granger-causes Hong Kong. However, this result is doubtful, because of serious serial correlation at lag 1 in this relation. Increasing the lag order to 2 resolves the serial correlation problem and yields the same causality findings, although heteroskedasticity emerges now in the relation for Singapore. Oddly enough, at $l = 8$ also the relationship for Hong Kong becomes heteroskedastic, whereas serial correlation re-emerges for the Singapore equation.

In Table 10 trivariate VAR results are given for Hong Kong, Singapore and Korea, for which in Table 6 no evidence of cointegration could be established. Both AIC and SIC suggest the lag order to be 0 again, which is clearly misleading, given the significance of t and F tests. At $l = 1$, two causality relationships seem detected, namely that Singapore Granger-causes both Hong Kong and Korea. However, both equations show significant first-order serial correlation, so before drawing any conclusions first the lag order of the VAR model should be increased. For lag order 2 the causality relationships are still the same, but BP becomes significant, so at least heteroskedasticity robust t and F statistics should be used now. At lag order 8 again some serial correlation diagnostics become significant (as in Table 9), whereas now Singapore seems to be Granger-caused by both Hong Kong and Korea too. If the model for $l = 2$ were correctly specified, one would not expect to find significant serial correlation in the $l = 8$ model. So, a plausible explanation for this is, that the actual data generating process for the revenues at these markets is in fact not nested in any of the VAR models of the Tables 9 and 10, simply because unintentionally particular important determining factors have been omitted. Apparently in the $l = 2$ models the serial correlation tests lack power to detect these omissions.

Table 11 reports results of pentavariate VAR models for all five markets. BP and JB are always significant at the various lag orders; we removed their p-values from the table in order to save space. SIC and AIC favor the lag orders 0 and 2 respectively.

Table 9: P-values of Granger-causality tests and single equation diagnostics in VAR(l) models for the revenues of the stock indices of Hong Kong and Singapore.

		— HK —			— SG —					— order 2 —			— order 10 —			– Min –	
l	Dep	F	t	lag	F	t	lag	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
1	HK	0.90	0.90	1	0.03	0.03	2	0.35	0.00	0.61	0.06	0.06	0.97	0.61	0.61	0.02	1
	SG	0.67	0.67	1	0.85	0.85	1	0.29	0.00	0.69	0.63	0.63	0.27	0.26	0.26	0.14	7
2	HK	0.85	0.55	2	0.01	0.03	2	0.20	0.00	1.00	0.97	0.97	0.98	0.98	0.98	0.82	1
	SG	0.85	0.61	1	0.69	0.40	2	0.00	0.00	1.00	0.91	0.91	0.34	0.30	0.31	0.17	7
8	HK	0.92	0.27	8	0.11	0.04	1	0.00	0.00	1.00	0.56	0.56	1.00	0.56	0.58	0.14	5
	SG	0.86	0.13	7	0.20	0.05	6	0.00	0.00	1.00	0.58	0.58	1.00	0.04	0.04	0.02	7

Table 10: P-values of Granger-causality tests and single equation diagnostics in VAR(l) models for the revenues of the stock indices of Hong Kong, Singapore and.

		— HK —			— SG —			— KR —					— order 2 —			— order 10 —			– Min –	
l	Dep	F	t	lag	F	t	lag	F	t	lag	BP	JB	LB	LM	LMF	LB	LM	LMF	LMF	lag
1	HK	0.78	0.78	1	0.02	0.02	1	0.30	0.30	1	0.24	0.00	0.65	0.13	0.13	0.96	0.73	0.74	0.05	1
	SG	0.39	0.39	1	0.61	0.61	1	0.21	0.21	1	0.29	0.00	0.72	0.70	0.70	0.26	0.24	0.24	0.14	7
	KR	0.52	0.52	1	0.02	0.02	1	0.32	0.32	1	0.31	0.00	0.53	0.08	0.08	0.47	0.16	0.17	0.03	1
2	HK	0.75	0.51	2	0.01	0.02	1	0.65	0.37	1	0.00	0.00	1.00	0.96	0.96	0.98	0.98	0.98	0.79	1
	SG	0.63	0.36	1	0.68	0.45	2	0.49	0.23	1	0.00	0.00	1.00	0.89	0.89	0.32	0.26	0.26	0.15	7
	KR	0.71	0.44	1	0.02	0.03	1	0.68	0.40	1	0.00	0.00	0.97	0.17	0.17	0.55	0.33	0.34	0.17	2
8	HK	0.95	0.30	7	0.11	0.03	1	0.89	0.34	4	0.00	0.00	1.00	0.14	0.15	1.00	0.33	0.35	0.09	1
	SG	0.53	0.02	7	0.30	0.03	6	0.36	0.01	7	0.00	0.00	0.97	0.10	0.10	1.00	0.02	0.03	0.01	6
	KR	0.85	0.28	5	0.04	0.01	5	0.55	0.12	7	0.00	0.00	1.00	0.68	0.69	1.00	0.05	0.05	0.03	8

Table 11: P-values of Granger-causality tests and single equation diagnostics in VAR(l) models for the revenues of the stock indices of Hong Kong, Singapore, Korea, Shanghai and US.

		— HK —			— SG —			— KR —			— SH —			— US —			— order 2 —			— order 10 —			– Min –	
l	Dep	F	t	lag	F	t	lag	F	t	lag	F	t	lag	F	t	lag	LB	LM	LMF	LB	LM	LMF	LMF	lag
1	HK	0.80	0.80	1	0.02	0.02	1	0.30	0.30	1	0.24	0.24	1	0.82	0.82	1	0.62	0.24	0.24	0.96	0.85	0.85	0.09	1
	SG	0.30	0.30	1	0.54	0.54	1	0.25	0.25	1	0.82	0.82	1	0.31	0.31	1	0.80	0.18	0.18	0.19	0.09	0.09	0.04	6
	KR	0.55	0.55	1	0.02	0.02	1	0.32	0.32	1	0.85	0.85	1	0.82	0.82	1	0.52	0.11	0.11	0.46	0.20	0.20	0.04	1
	SH	0.20	0.20	1	0.67	0.67	1	0.51	0.51	1	0.51	0.51	1	0.30	0.30	1	0.94	0.76	0.76	0.30	0.23	0.23	0.23	10
	US	0.24	0.24	1	0.00	0.00	1	0.14	0.14	1	0.98	0.98	1	0.00	0.00	1	0.27	0.00	0.00	0.27	0.02	0.02	0.00	1
2	HK	0.99	0.86	1	0.00	0.02	1	0.49	0.31	1	0.24	0.23	2	0.01	0.00	2	0.98	0.28	0.29	0.98	0.85	0.86	0.29	2
	SG	0.60	0.36	1	0.55	0.31	2	0.44	0.22	1	0.86	0.61	2	0.00	0.00	2	0.96	0.01	0.01	0.25	0.02	0.03	0.01	2
	KR	0.65	0.49	2	0.01	0.03	1	0.62	0.32	1	0.94	0.75	2	0.05	0.02	2	0.96	0.16	0.16	0.45	0.24	0.25	0.16	2
	SH	0.12	0.10	2	0.47	0.25	2	0.72	0.50	1	0.78	0.48	1	0.61	0.33	1	1.00	0.43	0.43	0.31	0.18	0.19	0.19	10
	US	0.41	0.20	1	0.00	0.00	1	0.00	0.00	2	0.96	0.78	2	0.00	0.00	1	1.00	1.00	1.00	0.46	0.45	0.46	0.21	7
8	HK	0.95	0.22	7	0.04	0.01	2	0.72	0.11	4	0.01	0.00	6	0.19	0.00	2	0.99	0.13	0.14	1.00	0.13	0.15	0.03	5
	SG	0.39	0.01	7	0.15	0.01	6	0.29	0.01	7	0.27	0.08	8	0.00	0.00	2	0.99	0.62	0.64	0.99	0.03	0.04	0.03	8
	KR	0.74	0.29	6	0.01	0.01	5	0.45	0.14	4	0.02	0.01	4	0.08	0.02	2	0.97	0.06	0.06	1.00	0.03	0.04	0.04	8
	SH	0.19	0.09	8	0.19	0.00	5	0.85	0.29	6	0.20	0.06	3	0.79	0.12	8	0.99	0.46	0.47	0.89	0.34	0.38	0.38	10
	US	0.40	0.04	7	0.01	0.00	1	0.01	0.00	2	0.14	0.00	6	0.00	0.00	1	1.00	0.62	0.63	1.00	0.27	0.30	0.29	8

At $l = 1$ three causality relationships are found, namely that Singapore Granger-causes Hong Kong, Korea and the US. However, the equations for Singapore, Korea and the US suffer from serial correlation at the 5% level. This suggests to increase the lag order. For $l = 2$ four additional causality relationships emerge, namely that the US Granger-causes Hong Kong, Singapore and Korea, and that Korea, like Singapore, Granger-causes the US. However, the relation for Singapore shows significant serial correlation. Even taking $l = 8$ does not resolve this problem, but in fact (again) aggrevates it. Moreover, the VAR(8) model suggests another seven causality relationships. Table 12 reports for the same five markets the results of separate single equation regressions in which those for the Asian markets now include the most recent realized revenue for the US. The most striking differences with Table 11 are that for $l = 1$, now the US seems to Granger-cause Hong Kong, Singapore and Korea, while the Granger-causality relationships from Singapore to Hong Kong and Korea become insignificant. This highlights that just mechanically coping with the time zone difference in VAR models is inappropriate and can have devastating effects on inference.

Focussing only on those relationships in Table 12 which are not rejected by any of the serial correlation tests, and giving much more weight to the p-value of an F test than to an individual t tests (especially when l is large and a significant t test is just found at a high lag order), one could with substantial reservation infer that: (i) Hong Kong seems Granger-caused by all four other markets; (ii) Korea seems Granger-caused by Singapore, Shanghai and the US; Shanghai seems Granger-caused by none of the four markets, because the significant individual t-values at lag 5 for Singapore and at lag 8 for Hong Kong could just be incidental; the US seems Granger-caused by Singapore and Korea (and possibly very mildly by Shanghai); whereas for Singapore we have not found a regression model from this limited data set which passes the serial correlation diagnostics.

However, in this study our purpose was not to make strong claims regarding the major determining factors of the indices and revenues at these five markets. Our primary aim was to illustrate that a few isolated estimates and tests may seem very suggestive regarding the significance of particular relationships, but only when such results pass a relevant battery of diagnostic tests and also withstand extensions of the model, by including more markets and deeper lags they deserve to be taken seriously. Note that Table 10 suggests that Korea does not Granger-cause Hong Kong, whereas Table 12, after allowing the US to play its role as well, indicates the contrary.

These illustrations of causality testing have not only indicated again that information criteria cannot be trusted in determining the optimal lag order structure of a VAR model, but also that they certainly cannot reveal whether a VAR model suffers from omitted relevant determining variables such as lagged variables from other markets. In both respects diagnostic tests seem much more useful, but they are certainly no panacea. The only trustworthy tests against omitted variables seem to be constructive tests which include and test the significance of the actual yet omitted variables concerned. The common practice of VAR causality testing of revenue series without much concern

Table 12: P-values of Granger-causality tests and diagnostics of single equation for the revenues of the stock indices of Hong Kong, Singapore, Korea, Shanghai and US.

		— HK —			— SG —			— KR —			— SH —			— US —			— order 2 —			— order 10 —			– Min –	
l	Dep	F	t	lag	F	t	lag	F	t	lag	F	t	lag	F	t	lag	LB	LM	LMF	LB	LM	LMF	LMF	lag
1	HK	0.88	0.88	1	0.19	0.19	1	0.12	0.12	1	0.23	0.23	1	0.00	0.00	1	0.03	0.00	0.00	0.55	0.15	0.15	0.00	2
	SG	0.41	0.41	1	0.58	0.58	1	0.09	0.09	1	0.83	0.83	1	0.00	0.00	1	0.06	0.03	0.03	0.02	0.01	0.01	0.00	6
	KR	0.53	0.53	1	0.16	0.16	1	0.17	0.17	1	0.78	0.78	1	0.00	0.00	1	0.02	0.00	0.00	0.08	0.02	0.02	0.00	2
	SH	0.13	0.13	1	0.59	0.59	1	0.57	0.57	1	0.54	0.54	1	0.91	0.91	1	0.96	0.85	0.85	0.20	0.15	0.15	0.15	10
	US	0.24	0.24	1	0.00	0.00	1	0.14	0.14	1	0.98	0.98	1	0.00	0.00	1	0.27	0.00	0.00	0.27	0.02	0.02	0.00	1
2	HK	0.62	0.34	2	0.07	0.04	2	0.07	0.09	2	0.21	0.14	1	0.00	0.00	1	0.16	0.00	0.00	0.86	0.01	0.01	0.00	2
	SG	0.66	0.49	1	0.59	0.41	1	0.13	0.10	1	0.96	0.83	2	0.00	0.00	1	0.27	0.00	0.00	0.09	0.00	0.00	0.00	2
	KR	0.96	0.77	1	0.13	0.09	2	0.24	0.13	1	0.84	0.56	2	0.00	0.00	1	0.53	0.00	0.00	0.53	0.05	0.05	0.00	2
	SH	0.12	0.10	2	0.44	0.22	2	0.73	0.50	1	0.76	0.48	1	0.61	0.33	2	1.00	0.90	0.90	0.21	0.14	0.15	0.15	10
	US	0.42	0.20	1	0.00	0.00	1	0.00	0.00	2	0.96	0.77	2	0.00	0.00	1	1.00	1.00	1.00	0.46	0.45	0.46	0.21	7
8	HK	0.93	0.22	8	0.03	0.00	2	0.17	0.05	2	0.02	0.03	4	0.00	0.00	1	1.00	0.99	0.99	1.00	0.94	0.95	0.90	7
	SG	0.76	0.09	7	0.01	0.01	5	0.10	0.01	7	0.24	0.05	3	0.00	0.00	1	0.98	0.26	0.27	1.00	0.05	0.06	0.04	6
	KR	0.86	0.24	6	0.00	0.01	2	0.26	0.16	7	0.01	0.01	4	0.00	0.00	1	0.97	0.16	0.18	1.00	0.82	0.84	0.09	1
	SH	0.14	0.05	8	0.14	0.01	5	0.74	0.22	6	0.17	0.05	8	0.99	0.45	6	1.00	0.81	0.82	0.88	0.07	0.08	0.08	10
	US	0.40	0.04	7	0.01	0.00	1	0.01	0.00	2	0.14	0.00	6	0.00	0.00	1	1.00	0.62	0.63	1.00	0.27	0.30	0.29	8

regarding possibly omitted markets can be much improved, as we showed, especially for the situation that different time zones are involved. In that case, instead of a system, separate carefully specified single equations should be subjected to causality tests.

4. Conclusion

Much too often practitioners produce (or uncritically cite published) empirical results for which too few (or improper) diagnostics have been evaluated that do support their findings. As a rule any statistical evidence is based on an extensive set of adopted model assumptions. Evidence supporting the likely validity of many of these assumptions can usually be provided by so-called diagnostic tests or by demonstrating that the maintained model is not rejected when tested against a less restrictive alternative model formulation. By scrutinizing a great number of peer reviewed published papers on analyzing the interlinkages between various stock markets we demonstrate that in this literature there exists a tendency to take any empirical findings serious irrespective by what methodology and statistical techniques they have been obtained, whereas employing unsound statistical methodology is ubiquitous. We plea for bringing methodological issues more to the forefront, and stopping citing empirical results which have clearly been obtained by methods that as a rule will lead to strongly biased results for which neither statistical accuracy nor significance can be assessed in a proper way. This widespread submissive uncritical citation attitude makes one wonder too what the intrinsic value is of citation based indices on the quality of academic journals, because in the niche of studies examined here it is obvious that citations are obtained simply because of the particular topic being studied and not primarily because the study has been a useful stepping stone for further and deeper understanding. In Table 13 we give an overview of the various observations made regarding methodological shortcomings, dividing them in five major categories and some subcategories.

By evaluations using empirical data we illustrate the serious effects of wrongly omitted explanatory variables, such as variables of other influential markets and significant higher order lagged variables. Such kind of omitted variables affect the results of cointegration tests and Granger-causality tests, which are often the basis for further analyses and inference. In addition, particular popular information criteria are shown to provide seriously misleading results for assessing the appropriate lag length of VAR models. It is demonstrated that it is of great importance to conduct thorough checks and diagnostic tests to make sure that the underlying assumptions of the regressions are not rejected, whereas by simulation it is shown that particular often used tests for serial correlation are deficient, whereas more adequate alternative versions have been developed.

Table 13: Overview of major econometric shortcomings in the reviewed studies

	A	A1	B	B1	C1	C2	C3	C4	C5	D	E
Aityan et al. (2010)	✓		✓								
Arshanapalli et al. (1995)	✓	✓	✓	✓		✓			✓		
Beirne et al. (2010)	✓	✓							✓		
Burdekin & Siklos (2012)	✓					✓	✓				
Cheng & Glascock (2005)	✓	✓	✓	✓		✓					✓
Chow et al. (2011)	✓		✓			✓				✓	✓
Chung & Liu (1994)	✓	✓			✓				✓		
Glick & Hutchison (2013)	✓	✓	✓	✓			✓				✓
Gosh et al. (1999)	✓	✓	✓	✓							
Groenewold et al. (2004)	✓										
Huang et al. (2000)	✓	✓	✓	✓							
Huygebaert & Wang (2010)	✓	✓	✓	✓				✓			
Jayasuriya (2011)	✓	✓			✓				✓		
Kenett et al. (2012)	✓		✓								
Li (2007)	✓	✓	✓	✓					✓		
Li (2012)	✓								✓		
Singh et al. (2010)	✓	✓	✓	✓		✓					✓
Wang (2014)	✓	✓	✓	✓				✓			
Yang et al. (2006)	✓	✓									
Zhang & Li (2014)	✓		✓								✓
Zhang et al. (2009)	✓		✓	✓					✓		
Zhu et al. (2004)	✓		✓	✓				✓			

A = insufficient checks for omitted variables
A1 = and in particular insufficient checks for coefficient nonconstancy
B = insufficient diagnostic checking
B1 = and in particular insufficient appropriate testing for error serial correlation
C1 = inappropriate use of tests for normality
C2 = inappropriate use of t-tests
C3 = inappropriate use of tests of correlation
C4 = using joint restrictions test, where single restrictions are required too
C5 = using only high-order serial correlation tests
D = neglecting identification problems
E = presenting alternative results with mutually conflicting maintained hypotheses

References

Aityan, S.G., A.K. Ivanov-Schitz and S.S. Izotov (2010) Time-shift asymmetric correlation analysis of global stock markets. *Journal of International Financial Markets, Institutions & Money* 20: 590-605.

Arshanapalli, B., J. Doukas and L.H.P. Lang (1995) Pre and post-October 1987 stock market linkages between U.S. and Asian markets. *Pacific-Basin Finance Journal* 3: 57-73.

Beirne, J., G. M. Caporale, M. Schulze-Ghattas and N. Spagnolo (2010) Global and

regional spillovers in emerging stock markets: A multivariate garch-in-mean analysis. *Emerging Markets Review* 11: 250-260.

Burdekin, R.C.K. and P.L. Siklos (2012) Enter the dragon: Interactions between Chinese, US and Asia-Pacific equity markets, 1995-2010. *Pacific-Basin Finance Journal* 20: 521-541.

Chen, Z., J.F. Kiviet and W. Huang (2015) On the integration of China's main stock exchange with the international financial market. Mimeo.

Chow, G.C. and C.C. Lawler (2003) A time series analysis of Shanghai and New York stock price indices. *Annals of Economics and Finance* 4: 17-35.

Chow, G.C., C. Liu, and L. Niu (2011) Co-movements of Shanghai and New York stock prices by time-varying regressions. *Journal of Comparative Economics* 39: 577-583.

Chung, P.J. and D.J. Liu (1994) Common stochastic trends in pacific rim stock markets. *The Quarterly Review of Economics and Finance* 34: 241-259.

Edgerton, D. and G. Shukur (1999) Testing autocorrelation in a system perspective. *Econometric Reviews* 18: 343-386.

Glick, R. and M. Hutchison (2013) China's financial linkages with asia and the global financial crisis. *Journal of International Money and Finance* 39: 186-206.

Gosh, A., R. Saidi, and K.H. Johnson (1999) Who moves the Asia-Pacific stock markets–US or Japan? Empirical evidence based on the theory of cointegration. *The Financial Review* 34: 159-170.

Groenewold, N., S.H.K. Tang, and Y. Wu (2004) The dynamic interrelationships between the greater China share markets. *China Economic Review* 15: 45-62.

Huang, B.-N., C.-W. Yang, and J. W.-S. Hu (2000) Causality and cointegration of stock markets among the United States, Japan and the South China growth triangle. *International Review of Financial Analysis* 9: 281-297.

Huyghebaert, N. and L. Wang (2010) The co-movement of stock markets in East Asia; Did the 1997-1998 Asian financial crisis really strengthen stock market integration? *China Economic Review* 21: 98-112.

Jayasuriya, S. A. (2011) Stock market correlations between china and its emerging market neighbors. *Emerging Markets Review* 12: 418-431.

Kenett, D.Y., M. Raddant, T. Lux, and E. Ben-Jacob (2012) Evolvement of uniformity and volatility in the stressed global financial village. PLoS ONE 7(2): e31144.

Koundouri, P., N. Kourogenis, and N. Pittis (2016) Statistical modeling of stock returns: explanatory or descriptive? A historical survey with some methodological reflections. *Journal of Economic Surveys* 30: 149-164.

Li, H. (2007) International linkages of the Chinese stock exchanges: A multivariate GARCH analysis. *Applied Financial Economics* 17: 285-297.

Li, H. (2012) The impact of China's stock market reforms on its international stock market linkages. *The Quarterly Review of Economics and Finance* 52: 358-368.

Mahbubani, K. (2008) *The New Asian Hemisphere; The Irresistible Shift of Global Power to the East.* Public Affairs. New York.

Singh, P., B. Kumar, and A. Pandey (2010) Price and volatility spillovers across North American, European and Asian stock markets. *International Review of Financial Analysis* 19: 55-64.

Wang, L. (2014) Who moves East Asian stock markets? The role of the 2007-2009 global financial crisis. *Journal of International Financial Markets, Institutions & Money* 28: 182-203.

Yang, J., C. Hsiao, Q. Li, and Z. Wang (2006) The emerging market crisis and stock market linkages: further evidence. *Journal of Applied Econometrics* 21: 727-744.

Zhang, B. and X.-M. Li (2014) Has there been any change in the comovement between the chinese and US stock markets? *International Review of Economics and Finance* 29: 525-536.

Zhang, S., I. Paya, and D. Peel (2009) Linkages between Shanghai and Hong Kong stock indices. *Applied Financial Economics* 19: 1847-1857.

Zhu, H., Z. Lu, S. Wang, and A.S. Soofi (2004) Causal linkages among Shanghai, Shenzhen, and Hong Kong stock markets. *International Journal of Theoretical and Applied Finance* 7: 135-149.

Appendices

A. On omitted regressors

Suppose a relationship is examined for which a simple multiple regression model would be adequate and ordinary least-squares techniques would provide valid inferences (at least in samples of substantial size). What would be the consequences if some of the regressors were omitted from the regression? To answer this question it suffices to consider the simple two regressor model

$$y_t = \beta x_t + \gamma w_t + \varepsilon_t,$$

where x_t and w_t are scalar variables that have been observed together with the dependent variable y_t for observations $t = 1, ..., n$, whereas all three have been taken in deviation from their sample average, which permits to remove the intercept term from the regression. Extension to the case where both x_t and w_t are in fact vectors of variables, possibly containing various sequences of lagged variables as in VAR models as used in Granger-causality tests, are straight-forward, though mathematically more involved, and therefore avoided here.

We focus on the case where the data are time-series. Let $x^t = (x_1 \cdots x_t)'$ and $w^t = (w_1 \cdots w_t)'$. Appropriate interpretability of standard least-squares analysis requires validity of the three assumptions: (1) predeterminedness of all regressors, that is $E(\varepsilon_t \mid x^t, w^t) = 0$; (2) serial uncorrelatedness of the disturbances, that is $E(\varepsilon_t \varepsilon_s) = 0$ for $t \neq s$; and (3) homoskedasticity of the disturbances, or $E(\varepsilon_t^2) = \sigma^2$. Violation of (1) leads to biased estimates, irrespective of the size of the sample, and so does violation of (2) in VAR models. A major consequence of violation of (2) or (3) is that the standard

estimator for the variance of the coefficient estimates will be biased, which renders inferences based on standard test statistics and confidence intervals misleading. If the sample is not too small normality of the distribution of the disturbances is not essential. Of course, in practice validity of the three assumptions should always be verified as far as is possible. When interpreting standard least-squares inference on β and γ these three assumptions establish the so-called *maintained hypotheses*, which are required to be plainly valid.

Now suppose that a researcher deliberately or unknowingly omits regressor w_t. Then least-squares yields for β the estimator

$$\hat{\beta} = \Sigma_t x_t y_t / \Sigma_t x_t^2 = \beta + \gamma \Sigma_t x_t w_t / \Sigma_t x_t^2 + \Sigma_t x_t \varepsilon_t / \Sigma_t x_t^2.$$

Here the third term is the unavoidable random estimation error which varies around zero and decreases in magnitude with sample size. The second term represents a systematic deviation of $\hat{\beta}$ from β. This has a magnitude which does not decrease with the size of the sample; it does depend on four autonomous factors, since

$$\gamma \Sigma_t x_t w_t / \Sigma_t x_t^2 = \gamma r_{x,w} s_w \frac{1}{s_x},$$

where $r_{x,w}$ is the sample correlation between variables x_t and w_t, s_w is the sample standard deviation of the observations w_t, and similar for s_x. Hence, this systematic error (the inconsistency of $\hat{\beta}$) is small provided the product of these four factors is small. So, in case some of them are large, the remaining ones should be sufficiently small. If this is the case, estimator $\hat{\beta}$ will as a rule not differ much from its estimate obtained without omission of regressor w_t. However, even if that occurs, omission will nevertheless be harmful for least-squares inference in case the disturbances of the model omitting w_t (these are now represented by $\gamma w_t + \varepsilon_t$) do not obey the three fundamental conditions. It should be obvious that in case $\gamma \neq 0$ and w_t represents a series which is correlated with x_t then the predeterminedness assumption does no longer hold (but the consequences may be mild if γ and/or s_w/s_x are small). If $\gamma \neq 0$ and the w_t series exposes heteroskedasticity and/or serial correlation then the other two conditions may not hold in the model that through omission of w_t incorrectly imposes $\gamma = 0$. In that case inference on β will be unreliable because the required maintained hypotheses are not all satisfied. If it is just condition (3) that does not hold, it is possible to repair the inaccuracies of least-squares inference (make them robust to heteroskedasticity). In VAR-type models this is generally not possible when condition (2) is not fulfilled.

Hence, the model omitting w_t is only really superior to the model including w_t if $\gamma = 0$ and omission is fully justified. Otherwise, when $\gamma \neq 0$ and the omitted variable w_t is correlated with x_t (for instance, because it is just the one-period lag of smooth series x_t) or if omission does not lead to substantial bias, but the omitted variable is itself not time-independent, then proper interpretation of least-squares inference from the model wrongly imposing $\gamma = 0$ is impossible. This situation should be prevented by always carefully verifying whether the three maintained hypotheses on which least-squares inference is built do actually hold (see Appendix B). If they do not, the effect

of x_t on y_t can only be assessed either after a proper re-specification of the model, or by using an alternative for standard least-squares estimation.

B. On diagnostic testing

Diagnostic testing refers to all attempts made to verify validity of the *maintained hypotheses*. By these we indicate here all statistical aspects of the model that have to hold in order to be able to accurately interpret all inference techniques that one intends to employ regarding its parameter values. These statistical aspects should establish at least: (a) consistency of the estimators (estimators that are right on target in infinitely large samples), (b) confidence intervals for the unknown coefficient values with an actual coverage probability which converges for increasing sample size to the chosen nominal level (often 95%), and (c) actual significance levels (type I error probabilities) of tests on coefficient values which converge to their chosen nominal level of $100 \times \alpha\%$. When one chooses to analyze a linear model by standard least-squares techniques the maintained hypotheses are the three stated in Appendix A.

From Appendix A it should be clear that when in a linear model that has been estimated by ordinary least-squares one finds significant test outcomes for heteroskedasticity or for serial correlation the proper diagnosis might be that in fact improper restrictions have been imposed on regression coefficients. Hence, this may not call for some form of weighted least-squares or autoregressive transformation of the model, but for finding the omitted regressors. Candidates are not only really additional new variables, but also transformations of already included variables, such as lags, logs, squares, cross-products (interaction terms), dummy variables and cross-products with dummy variables (to represent variation in reaction coefficients over sub-samples).

Note that the predeterminedness assumption of a regressor may be violated due to an omitted regressor with which it happens to be correlated. This is resolved by including the omitted regressor. On the other hand, a regressor may be endogenous instead of predetermined because of reverse causality. This occurs when y_t is determined by x_t, while y_t itself is also one of the explanatory variables of x_t. Then x_t, since it depends on y_t, must also be correlated with ε_t. In that case variables y_t and x_t are called jointly dependent and valid inference on β (the effect of x_t on y_t) cannot be obtained by ordinary least-squares. However, provided the equation for y_t has sufficient unique characteristics (usually in the form of omitted variables whose valid omission is beyond doubt) which guarantee identification (see Appendix D), valid inference can be obtained by using the correctly omitted regressors as so-called external instrumental variables in two-stage least-squares estimation. For the latter the predeterminedness assumption on x_t in the maintained hypotheses has to be replaced by so-called exclusion restrictions (correctly omitted regressors). So, the moral is here: wrongly omitted regressors are as a rule harmful for inference, whereas rightly omitted regressors are obligate for valid inference when reverse causality may be at play. In the latter case the precision of inference will be poor when x_t depends only weakly on the instrumental variables.

So, useful diagnostic tests are tests for heteroskedasticity (standard software provides tests against various types of heteroskedasticity), tests for serial correlation of suitable order (the popular Durbin-Watson test just checks first-order serial correlation and presupposes that all regressors are strictly exogenous, so there should be no lagged-dependent variables among the regressors as in VAR models), tests for the significance of omitted relevant variables (many implementations are possible, also covering tests for structural changes and for nonlinear dependencies). Also appropriate tests for serial correlation and tests for reverse causality take the form of omitted regressor tests because they boil down to testing the significance of particular (transformations of) residuals when added to the original model. Only when a model specification has been found which passes a battery of diagnostic tests (all yielding p-values preferably well above 0.10, or at least above 0.05), then building on the maintained hypotheses implied by the chosen estimation technique seems warranted. This finally opens the door to the phase in which inference on the coefficients of the model can be produced and interpreted with reasonable degree of trust.

C. On significance tests of simple correlations

As is well-known a test on the significance of β in the simple linear regression model $y_t = \alpha + \beta x_t + \varepsilon_t$ $(t = 1, ..., n)$ is obtained by the least-squares based test statistic $t = \hat{\beta}/se(\hat{\beta})$, where $\hat{\beta} = \Sigma_t \tilde{x}_t \tilde{y}_t / \Sigma_t \tilde{x}_t^2$ and $se(\hat{\beta}) = [(n-2)^{-1} \Sigma_t e_t^2 / \Sigma_t \tilde{x}_t^2]^{1/2}$. Here $\tilde{x}_t = x_t - \bar{x}$ with $\bar{x} = n^{-1} \Sigma_t x_t$ (and likewise for $\tilde{y}_t$) and $e_t = y_t - \hat{\alpha} - \hat{\beta} x_t = \tilde{y}_t - \hat{\beta} \tilde{x}_t$, because $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$. Some simple algebraic manipulations show that $t = \hat{\rho} / [(1 - \hat{\rho}^2)/(n-2)]^{1/2}$, where $\hat{\rho} = \Sigma_t \tilde{x}_t \tilde{y}_t / [\Sigma_t \tilde{x}_t^2 \Sigma_t \tilde{y}_t^2]^{1/2}$ is the simple sample correlation coefficient of the series $\{y_t, x_t; t = 1, ..., n\}$.

Therefore, testing whether or not the correlation $\rho = Cov(y_t, x_t) / [Var(x_t) \cdot Var(y_t)]^{1/2}$ is zero by reference to the statistic t and a critical value of the Student distribution with $n - 2$ degrees of freedom requires validity of similar maintained hypotheses as testing the significance of β in the simple linear regression model. Sufficient for that is $\varepsilon_t \mid x^t \sim IIN(0, \sigma^2)$, which represents the three maintained hypotheses mentioned in Appendix A plus normality of the disturbances. In samples of reasonable size the normality is of minor importance. However, for asymptotic validity of the test it is essential that subtracting from y_t its conditional expectation expressed linearly in just the single variable x_t should remove all its time-dependence and heteroskedasticity and thus yield random white-noise.

Hence, a test on the significance of a simple correlation coefficient can only be credibly and usefully interpreted when supplemented at least by insignificant test statistics on both (higher-order) serial correlation and heteroskedasticity of the disturbances of the corresponding simple linear regression, and preferably also some further evidence on constancy of the coefficients α and β and absence of the devastating effects of possibly omitted explanatory variables. Thus, if a VAR specification for y_t seems credible, then performing the above test on correlation (which assumes a static regression) does not

make sense.

D. On lack of identification

Consider a system, for which scalar variables $y_t^{(1)}$, $y_t^{(2)}$ and x_t $(t = 1, ..., n)$ have been observed for a sample of size n, that is given by the two equations

$$y_t^{(1)} = \beta_1 y_t^{(2)} + \gamma_1 x_t + \varepsilon_t^{(1)},$$
$$y_t^{(2)} = \beta_2 y_t^{(1)} + \gamma_2 x_t + \varepsilon_t^{(2)}.$$

Here the unobserved variables $\varepsilon_t^{(j)}$ $(j = 1, 2)$ are disturbance terms, and x_t is in both equations an explanatory variable with unknown coefficient values γ_1 and γ_2. The variables $y_t^{(j)}$ $(j = 1, 2)$ are characterized by reverse causality, provided both β_1 and β_2 are nonzero. To keep things simple, we assume that both disturbances are white-noise series, which are not necessarily mutually uncorrelated, and that x_t (which could easily be extended to a vector of control variables) is predetermined and thus uncorrelated with both current disturbance terms.

Without any further assumptions, for instance on the actual value of at least one of the four coefficients β_j and γ_j $(j = 1, 2)$, all these coefficients are unidentified. This expresses that, whatever method one uses to obtain estimates $\hat{\beta}_j$ and $\hat{\gamma}_j$ $(j = 1, 2)$, it is impossible to find out whether these refer to the above system parameters or to some hidden linear transformations of them.

This can be clarified as follows. Let c_1 and c_2 be arbitrary real scalars. Now consider the rescaled system

$$c_1 y_t^{(1)} = c_1\beta_1 y_t^{(2)} + c_1\gamma_1 x_t + c_1\varepsilon_t^{(1)},$$
$$c_2 y_t^{(2)} = c_2\beta_2 y_t^{(1)} + c_2\gamma_2 x_t + c_2\varepsilon_t^{(2)}.$$

By addition of these two equations we obtain

$$(c_1 - c_2\beta_2)y_t^{(1)} = (c_1\beta_1 - c_2)y_t^{(2)} + (c_1\gamma_1 + c_2\gamma_2)x_t + (c_1\varepsilon_t^{(1)} + c_2\varepsilon_t^{(2)}).$$

From this we can again establish a two-equation system by normalizing the coefficients of $y_t^{(1)}$ and $y_t^{(2)}$ respectively, giving

$$y_t^{(1)} = \frac{c_1\beta_1 - c_2}{c_1 - c_2\beta_2} y_t^{(2)} + \frac{c_1\gamma_1 + c_2\gamma_2}{c_1 - c_2\beta_2} x_t + \frac{c_1\varepsilon_t^{(1)} + c_2\varepsilon_t^{(2)}}{c_1 - c_2\beta_2},$$
$$y_t^{(2)} = \frac{c_2\beta_2 - c_1}{c_2 - c_1\beta_1} y_t^{(1)} + \frac{c_1\gamma_1 + c_2\gamma_2}{c_2 - c_1\beta_1} x_t + \frac{c_1\varepsilon_t^{(1)} + c_2\varepsilon_t^{(2)}}{c_2 - c_1\beta_1}.$$

Of course we assumed here that c_1 and c_2 are such that $c_1/c_2 \neq \beta_2$ and $c_2/c_1 \neq \beta_1$. Using obvious definitions for the new symbols introduced below, the latter system can also be expressed as

$$y_t^{(1)} = \beta_1^* y_t^{(2)} + \gamma_1^* x_t + u_t^{(1)},$$
$$y_t^{(2)} = \beta_2^* y_t^{(1)} + \gamma_2^* x_t + u_t^{(2)},$$

which demonstrates that it is indistinguishable from the initial one. This highlights that each of its equations cannot be identified from an arbitrary linear combination of the two equations after this has been scaled with respect to the same left-hand variable.

In fact, if one would use least-squares to estimate the equation with $y_t^{(1)}$ as the dependent variable, then one will obtain coefficient estimates which minimize the residual sum of squares. This corresponds to implicitly choosing c_1 and c_2 such that, if $Var(\varepsilon_t^{(1)}) = \sigma_1^2$, $Var(\varepsilon_t^{(2)}) = \sigma_2^2$ and $Cov(\varepsilon_t^{(1)}, \varepsilon_t^{(2)}) = \sigma_{12}$, the resulting

$$Var(u_t^{(1)}) = \frac{c_1^2\sigma_1^2 + 2c_1c_2\sigma_{12} + c_2^2\sigma_2^2}{(c_1 - c_2\beta_2)^2}$$

will be minimal.

E. Chicken and egg in testing

In Appendix B we warned against interpreting inference on coefficients of a model produced by estimators and test statistics before sufficient evidence has been produced by diagnostic tests demonstrating the credibility of all the maintained hypotheses. Now one may wonder whether these diagnostic tests themselves can safely be interpreted. They may require their own maintained hypotheses whose validity has not been verified yet. How to proceed?

Let the current specification of a relationship be represented by the simple model $y_t = \beta x_t + \varepsilon_t$ $(t = 1, ..., n)$, where for the sake of simplicity x_t and β are again scalars. The candidate set of maintained hypotheses for inference based on standard least-squares estimates is: $E(\varepsilon_t \mid x^t) = 0$, $E(\varepsilon_t\varepsilon_s) = 0$ for $t \neq s$, and $E(\varepsilon_t^2) = \sigma^2$. Now consider the extended model $y_t = \beta x_t + \gamma w_t + \varepsilon_t^*$, where $\varepsilon_t^* = \varepsilon_t - \gamma w_t$. Also w_t is taken as a scalar here, but it could easily be generalized to be a vector containing possibly omitted regressors (for instance dummy variables and interactions of the elements of x_t with those dummies, or squares of elements of x_t or really additional alternative variables) or w_t may contain constructs like lagged residuals or reduced form residuals as in particular diagnostic tests on serial correlation of disturbances and endogeneity of some of the regressors. Now, in case γ is scalar, the diagnostic test is based on the usual t-ratio for testing the null-hypothesis $\gamma = 0$ in the extended regression. Rejection or not of this hypothesis, which will be interpreted as rejection or not of the maintained hypotheses of the model with just the regressors x_t, is based on comparing this test statistic with a critical value of the t distribution. Why is this a sound procedure?

Note that (at least in large samples) the test statistic will be distributed according to the appropriate t distribution if the maintained hypotheses in the regression with just regressor x_t do hold indeed and additionally $E(\varepsilon_t \mid w_t) = 0$ which implies $\gamma = 0$. When finding a significant test statistic this may be either due to validity of the maintained hypotheses, though having obtained (with a small probability converging towards $100\alpha\%$) an unlikely extreme draw from the t distribution (this then leads to a type I error), or due to $E(\varepsilon_t \mid w_t) \neq 0$ and thus $\gamma \neq 0$. The latter in principle implies invalidity of the maintained hypotheses under test, because it seems very likely that

$E(\varepsilon_t \mid x_t) = E(\gamma w_t + \varepsilon_t^* \mid x_t) \neq 0$ when $E(w_t \mid x_t) \neq 0$. In addition, possibly $Var(w_t \mid x_t)$ is nonconstant and $Cov(w_t w_s \mid x_t) \neq 0$ for some $t \neq s$, which would also imply invalidity of the maintained hypotheses. Only in case $E(\varepsilon_t \mid w_t) \neq 0$ and thus $\gamma \neq 0$, whereas at the same time w_t is such that $E(w_t \mid x_t) = 0$, and $Var(w_t \mid x_t) = \sigma_w^2$ and also $Cov(w_t w_s \mid x_t) = 0$ for $t \neq s$, the test statistic will not follow the t distribution and thus may reject with probability exceeding α, although the maintained hypotheses could be valid. In these latter extra-ordinary circumstances (note that the three required characteristics of the moments of w_t can be verified from the observed data) γw_t is such that omitting it is relatively harmless, because it can fully be absorbed in the disturbances without affecting the maintained hypotheses. However, using the significant diagnostic to re-specify the model may in fact be preferable. Its maintained hypotheses will be valid too if those of the original model were valid.

Hence, we conclude that a significant diagnostic should always lead to re-specification of the model and a corresponding reformulation of the maintained hypotheses. On the other hand, very little can be concluded from an insignificant diagnostic. Of course, it could be a consequence of validity of the maintained hypotheses, but it could as well be due to lack of power of the particular diagnostic test. This will occur when (some elements of) the maintained hypotheses are invalid, but variable w_t embodies insufficiently the aspects that have been omitted from the current specification of the model for the analyzed actual data generating process. This, for instance, happens when a regressor v_t has wrongly been omitted, whereas the variable w_t used in the diagnostic test is hardly correlated with v_t.

Note that the type I error probability of diagnostic testing can be controlled, irrespective of the quality of the specifications of the model with and the one without the extra regressors w_t. This is because its null hypothesis (full validity of the maintained hypotheses in the model without w_t) involves all the assumptions that lead to the test statistic having its known null-distribution. When the test statistic is not significant this does not imply that these maintained hypotheses are valid, but if the diagnostic test statistic is significant this (apart from possible type I errors) clearly demonstrates invalidity of these maintained hypotheses, although it does not convey that the model extended with variables w_t implies valid reformulated maintained hypotheses.

What is the difference with a standard t test in the initial model of testing a null-hypothesis like $\beta = c$ (with c some real number)? The difference is that its usual interpretation when the test statistic has a significant value is simply $\beta \neq c$, whereas as long as the maintained hypotheses have not yet been verified, the interpretation should be that either $\beta \neq c$ and/or elements of the maintained hypotheses do not hold. The latter less comfortable addition may only be swept under the carpet if, prior to performing tests on β, extensive testing of the maintained hypotheses did not lead to its rejection.

F. On testing in an unidentified equation

Consider the simple system of two unidentified regression models

$$y_1 = X_1\beta_1 + \varepsilon_1 = y_2\beta_{11} + X_0\beta_{10} + \varepsilon_1,$$
$$y_2 = X_2\beta_2 + \varepsilon_2 = y_1\beta_{21} + X_0\beta_{20} + \varepsilon_2.$$

For $j \in \{1,2\}$ the $n \times 1$ vectors y_j and ε_j contain the observations of the two endogenous variables and the disturbance terms respectively, the $n \times K$ matrices X_j represent the regressor matrices, which have $K-1$ columns in common, namely X_0, and the $K \times 1$ vectors β_j denote the regression coefficients. Using the notation $P_j = X_j(X_j'X_j)^{-1}X_j'$ and $M_j = I - P_j$ for $j \in \{0,1,2\}$ in standard results for partitioned regression, the least-squares estimates for the scalar coefficients β_{11} and β_{21} can be expressed as

$$\hat{\beta}_{11} = y_2'M_0y_1/y_2'M_0y_2 \text{ and } \hat{\beta}_{21} = y_1'M_0y_2/y_1'M_0y_1.$$

For $j \in \{1,2\}$ the residual vectors of the two regressions are given by $e_j = M_jy_j$, their estimated disturbance variances by $s_j^2 = y_j'M_jy_j/(n-K)$ and the estimated variance of their first coefficients by $\widehat{Var}(\hat{\beta}_{11}) = s_1^2/y_2'M_0y_2$ and $\widehat{Var}(\hat{\beta}_{21}) = s_2^2/y_1'M_0y_1$. Now we find for the t-ratio for coefficient β_{11} the expression

$$\frac{\hat{\beta}_{11}}{se(\hat{\beta}_{11})} = \frac{y_2'M_0y_1/y_2'M_0y_2}{s_1/(y_2'M_0y_2)^{1/2}} = \frac{(n-K)^{1/2}y_2'M_0y_1}{(y_1'M_1y_1)^{1/2}(y_2'M_0y_2)^{1/2}}$$

and for that of β_{21}

$$\frac{\hat{\beta}_{21}}{se(\hat{\beta}_{21})} = \frac{(n-K)^{1/2}y_1'M_0y_2}{(y_2'M_2y_2)^{1/2}(y_1'M_0y_1)^{1/2}}.$$

Note that these two expressions have the same numerator. Next we will demonstrate that they also have the same denominator by using for the partitioned regressor matrices a result regarding the projection matrices P_j for $j \in \{1,2\}$, namely $P_1 = P_0 + P_{M_0y_2}$, which yields $M_1 = M_0 - P_{M_0y_2}$, where self-evidently $P_{M_0y_2} = M_0y_2(y_2'M_0y_2)^{-1}y_2'M_0$. Likewise $M_2 = M_0 - P_{M_0y_1}$. Substitution in the squared denominator for the t-ratio for β_{11} yields

$$y_1'M_1y_1y_2'M_0y_2 = y_1'[M_0 - P_{M_0y_2}]y_1y_2'M_0y_2 = y_1'M_0y_1y_2'M_0y_2 - (y_1'M_0y_2)^2$$

and the same is found for the other. Both test statistics are found to be equal to

$$\frac{\hat{\beta}_{j1}}{se(\hat{\beta}_{j1})} = \frac{\hat{\rho}_{12,0}}{[(1-\hat{\rho}_{12,0}^2)/(n-K)]^{1/2}}, \text{ where } \hat{\rho}_{12,0} = \frac{y_2'M_0y_1}{(y_1'M_0y_1)^{1/2}(y_2'M_0y_2)^{1/2}}.$$

Both test exactly the same hypothesis, namely whether (under a set of maintained hypotheses) the marginal correlation between the variables y_1 and y_2, after both variables have been cleared from their linear dependence on the variables in X_0, is zero or not. No separate conclusions can be drawn from the two equivalent t-ratio's.