Microeconometric Dynamic Panel Data Methods: Model Specification and Selection Issues

Kiviet, Jan

University of Amsterdam

5 April 2019

Microeconometric Dynamic Panel Data Methods: Model Specification and Selection Issues

Jan F. Kiviet[*]

Version of 5 April 2019

JEL Classifications: *C18, C23, C26, C52, C81, C87, J23*

Keywords: *classification of explanatories; dynamic impacts; interactions; feedback mechanisms; generalized method of moments; labor demand; model building strategy; short panels.*

Abstract

A motivated strategy is presented to find step by step an adequate model specification and a matching set of instrumental variables by applying the programming tools provided by the Stata package Xtabond2. The aim is to implement generalized method of moment techniques such that useful and reasonably accurate inferences are extracted from an observational panel data set on a single microeconometric structural presumably dynamic behavioral relationship. In the suggested specification search three comprehensive heavily interconnected goals are pursued, namely: (i) to include all the relevant appropriately transformed possibly lagged regressors, as well as any interactions between these if it is required to relax the otherwise very strict homogeneity restrictions on the dynamic impacts of the explanatories in standard linear panel data models; (ii) to correctly classify all regressors as either endogenous, predetermined or exogenous, as well as being either effect-stationary or effect-nonstationary, implying which internal variables could represent valid and relatively strong instruments; (iii) to enhance the accuracy of inference in finite samples by omitting irrelevant regressors and by profitably reducing the space spanned by the full set of available internal instruments. For the various tests which trigger the decisions to be made in the sequential selection process the relevant considerations are spelled out to interpret the magnitude of p-values. Also the complexities to establish and interpret the ultimately established dynamic impacts are explained. Finally the developed strategy is applied to a classic data set and is shown to yield new insights.

1. Introduction

When analyzing a panel data set covering a short time span for a relatively large number of subjects (families, companies, etc.), and especially when one may have to deal too with

[*] Professor of Econometrics, Amsterdam School of Economics, University of Amsterdam, PO Box 15867, 1001 NJ Amsterdam, The Netherlands (j.f.kiviet@uva.nl).

possible joint dependence of some of the variables occurring in the structural behavioral relationship under study, the best option seems to employ the generalized method of moments (GMM) estimation technique. Its use in this context for single dependent variables following a continuous distribution has been promoted especially by Arellano and Bond (1991). Its frequent application in practice has been stimulated in particular by the development of the Stata package Xtabond2 (StataCorp LLC, College Station, TX, USA) by Roodman (2009a). Below we will develop a strategy for searching an adequate specification when modeling dynamic panel data relationships using GMM. This strategy is largely in line with the one already set forth in Bond (2002), though here it is supplemented with various hints regarding particular additional methodological and practical concerns and with detailed instructions for employing Xtabond2.

In what follows, the reader is supposed to be already familiar with the technicalities of GMM for dynamic panel data models, including its one and two-step implementations as suggested by Arellano and Bond (1991) and Blundell and Bond (1998), and also with Sargan-Hansen tests for (subsets of) over-identification restrictions, tests for first and second order serial correlation as developed by Arellano and Bond (1991) and the various options to aim at improved finite sample performance, such as robustifying 1-step and correcting 2-step GMM standard errors and mitigating estimator bias by constraining the set of internal instrumental variables in one way or another. More details on and further useful references regarding most of these technicalities can be found in the first part of Kiviet, Pleus and Poldermans (2017). Its second part, which addresses major aspects of the actual performance under particular circumstances as obtained from simulation experiments of the various available individual statistical tools, will sometimes be referred to when discussing the vulnerability of the sequential stages of the model specification and instrument selection strategy to be developed here.

So, the major purpose of this study is to provide clues on how to use all the available inference tools and their various options effectively in a situation where an empirical panel data set is available to model one particular dependent variable, but where it is yet largely unknown which regressors should be included in an appropriate model and which variables seem proper instruments. Section 2 discusses in various subsections general methodological considerations regarding the confrontation of scientific aspirations and practical limitations in the context of analyzing panel data. Special attention will be paid to the role played by p-values, which are being substantiated in Appendix A. Section 3 discusses the consequences for dynamic reaction patterns of including in the model lags of the dependent and of other variables as regressors. Special attention is being paid to the complex dynamic impacts of regressors affected themselves by feedbacks, and also to how the standard linear dynamic panel data model can be augmented by interaction terms in order to allow for heterogeneity in dynamic impact multipliers. Next, Section 4 focuses on some further practical considerations typical when analyzing dynamic panel models and employing Xtabond2. Also, from all foregoing considerations, it compiles a comprehensive practical 10-stages search strategy. This aims to find from an appropriate panel data set an adequate actual empirical model specification with a matching GMM implementation and sensible inferences on the major characteristics of the established relationship. Section 5 employs the developed strategy to the classic data set also used in Arellano and Bond (1991) and Blundell and Bond (1998). Section 6 draws some general conclusions.

2. General methodological considerations

The econometric methodology that we will present aims at obtaining for a single structural (i.e. causal) relationship estimates of the unknown parameters which are consistent and relatively efficient as well as reasonably accurate. So, the coefficient estimates should converge fairly fast to their true values for a sample with a hypothetically increasing number of independently drawn subjects, but in the actually available finite sample they should also have moderate bias and standard deviation. To achieve this, one should arrive, after several stages to be described below, at a parsimonious though adequate specification of the explanatory part of the model with an effective set of instrumental variables. This set should contain at least as many instrumental variables as the model has unknown coefficients to satisfy the necessary order condition for accomplishing identification. For consistency the instrumental variables should be valid (uncorrelated with the unexplained idiosyncratic disturbances of the estimated model). Efficiency is fostered by pursuing on the one hand the effectivity of the instruments (they should have substantial correlation in a multivariate sense with the explanatory variables of the model) and on the other hand the parsimony of the model specification (exclude redundant explanatories). The two objectives consistency and efficiency are asymptotic in nature and are therefore rather abstract. It is mostly taken for granted –although below we will encounter particular contrarian regularities– that by pursuing these abstract asymptotic goals the actual accuracy in the sample at hand will benefit as well.

This section has six subsections. In the first we argue in general terms how to start-up a specification search by clarifying various hazards and opportunities. In the second subsection most of these aspects are formalized. Then, against this formal background, the third subsection discusses how to obtain an initial general statistically adequate model as a yet rough but precious diamond, from which through a careful sequential grinding process the pursued valuable object should finally emerge. The fourth discusses the major dilemmas the practitioner faces in the various stages of such a specification search. The fifth subsection highlights the crucial role plaid by the interpretation of p-values in the sequential decision processes to be laid out. And the sixth and final subsection shortly addresses the major consequences of using the same sample data in sequences of hypothesis tests.

2.1. Getting started

To foster in a practical modelling situation the goals of consistency and efficiency it seems a good idea to start off with a model specification which does not impose too many implicit or explicit coefficient restrictions, but does limit in a particular sense at the same time the number of exploited moment conditions. Regarding the coefficients this implies not starting with an overly simplistic model specification but including in the initial specification any variables (in as far as the available data set permits) which may possibly be relevant for the direct determination of the current dependent variable according to the available substantive economic theory. Because most empirical relationships are dynamic, statistical adequacy of this model specification may most probably require (see Section 3) the inclusion of one or more lags of each individual explanatory variable, as well as lags of the dependent variable. At the same time, inclusion of sufficient lagged variables may prevent occurrence of serial correlation of the error terms. The reason for this recommended lavishness regarding the inclusion of (lagged) regressors is that when relevant explanatory variables would be wrongly omitted, this has devastating consequences for the error term of the misspecified model, which will impede finding valid instruments and realize consistency of the estimators.

Regarding the candidate instrumental variables, the strategy to adopt initially a limited number of moment conditions should imply the following: To account for the possible relevance of reverse causality initially all current (i.e. unlagged) regressors should preferably be treated as endogenous instead of predetermined or exogenous. Regressors should be classified immediately as exogenous only in case it is obvious that they obtain their values completely outside the mechanism under study, such as time-dummies, hours of sunshine or current age of the subject. In this way one abstains as much as possible from using any instrumental variables and corresponding moment conditions which could possibly be invalid.

In the present context of dynamic structural panel data models to be estimated by GMM all instrumental variables usually stem from possibly linearly transformed lags (or leads or current observations) of the explanatory variables. Direct feedback from the current dependent variable –which includes the current error term– into a current explanatory variable renders the latter endogenous. Presence of lagged feedback of the dependent variable into an explanatory variable –in the absence of direct feedback– renders it predetermined. Occurrence of any possibly delayed feedbacks, together with specific assumptions on any form of serial correlation of the disturbance term, determine the suitability of an explanatory variable (or any of its leads or lags) as a valid instrumental variable. In case of non-zero first- and higher-order serial correlation of the error terms, most or even all (lagged) variables affected by feedbacks become invalid instruments. To avoid this, dynamic panel relationships should preferably be specified such that it is reasonable to assume that all dynamics has been accounted for by including sufficiently lagged regressors in the explanatory part of the model, resulting in a remaining idiosyncratic error term that –when supported by appropriate empirical evidence– can be assumed to be serially uncorrelated.

Usually the equation which is actually estimated will be taken in first differences, in order to eliminate unobserved time-constant heterogeneity represented by the so-called individual effects. Then, when the errors of the level equation are serially uncorrelated indeed, those of the first-differenced equation have negative first-order serial correlation of moving average form, with a first-order serial correlation coefficient -0.5 and zero second and higher-order serial correlation coefficients. In that case, the first difference of the error term will be uncorrelated with the second and higher order lags of any explanatory variable, irrespective of any instantaneous or lagged feedbacks from the dependent variable in that variable. When a regressor is affected just by lagged feedback from the dependent variable then the first lag of such a regressor must be uncorrelated with the first difference of serially uncorrelated disturbances, whereas in case of no feedbacks (i.e. for a strictly exogenous regressor) any lag or lead of that variable will be a valid instrument with respect to the error term, irrespective of its transformation or its serial correlation. Hence, from an exogenous regressor many more instruments can be constructed than from a predetermined regressor, whereas from an endogenous regressor fewer instruments can be constructed than from a predetermined one.

From the above informal introduction we can already distil three prominent guidelines which will be more formally substantiated later in this study:

(i) Limiting the number of employed orthogonality conditions in the initial stage of the search process should be understood as that one should restrain oneself in the first stages of the modelling process in using current and first lagged regressors as instruments for the first-differenced equation, except when it is obvious that a regressor is strictly exogenous.

(ii) In general one can use at least as many instruments constructed from an exogenous regressor as one can use from a predetermined regressor, and similarly at least as many from a predetermined regressor as one derives from an endogenous one. This is in contrast with what most practitioners do these days. In dynamic panel research, often the habit from standard static simultaneous equation estimation is followed, where each exogenous regressor is just instrumented by itself. However, in general, lags of exogenous regressors will establish strong and valid instruments for any non-exogenous regressors, in particular for regressors affected by immediate or lagged feedbacks from the dependent variable, in particular the lagged dependent regressor variables themselves.

(iii) Under the assumption of serially uncorrelated errors any second or higher order lagged regressor should constitute a valid instrument for the first-differenced equation. Hence, when upon testing the validity of such instruments a rejection results, the conclusion should not be that they are simply invalid and should be removed from the set of instruments. On the contrary, the conclusion should be that apparently the model is misspecified. Because, when a lagged regressor has wrongly been omitted from the model, the implied error term of this misspecified model will usually be correlated with that omitted variable. Self-evidently, this lagged variable and higher order lags of it with which it is most probably autocorrelated, should then turn out to be invalid instruments for the misspecified model. So, although counter-intuitively at first view, the proper reaction to such instrument invalidity of an at least twice lagged regressor should not be to remove it form the set of instruments, but to keep it as an instrument and use it as a regressor as well to force more appropriate disturbances. Possibly, however, other variables that happen to be correlated with these "invalid instruments" have been omitted from the regression. Anyhow, if at least twice lagged regressors turn out to be invalid instruments this implies that the regression equation has not yet been specified adequately.

2.2. Formalization of some of the principles just set forth

Following the above considerations an initial model specification should have the general form

$$y_{i,t} = \sum_{l=1}^{p_0} \gamma_l y_{i,t-l} + \sum_{m=1}^{M} \sum_{l=0}^{p_m} \beta_l^{(m)} x_{i,t-l}^{(m)} + \sum_{s=2}^{T} \tau_s d_{i,t}^{(s)} + \eta_i + \varepsilon_{i,t}, \tag{2.1}$$

where $i = 1, ..., N$ and $t = 1, ..., T$ and the single dependent variable $y_{i,t}$ is explained by: (i) $p_0 \geq 1$ lags of this dependent variable, (ii) $p_m \geq 0$ lags of M distinct explanatory variables $x_{i,t}^{(m)}$, (iii) $T - 1$ time-dummies, where $d_{i,t}^{(s)} = 1$ for $t = s$ and zero otherwise, (iv) N random or fixed individual effects η_i, and (v) idiosyncratic disturbances $\varepsilon_{i,t}$. This family of general linear (in its coefficients) autoregressive distributed lag panel data models has some very specific family members, namely: (a) when $p_m = 0$ for all $1 \leq m \leq M$ and $p_0 = 1$ with $|\gamma_1| < 1$ we have a so-called partial adjustment panel model; (b) this specializes to a static panel model when $\gamma_1 = 0$; (c) and otherwise, when

$p_0 = 1$ with $\gamma_1 = 0$, while $0 \leq p_m \ll \infty$ for $1 \leq m \leq M$ with some $p_m > 0$, the model is known as a classic finite distributed lag panel data model. In many empirical cases, especially for annual data, the orders $p_0, ..., p_M$ will probably not exceed 1 or 2.

We assume that all individual variables $y_{i,t}$ and $x_{i,t}^{(m)}$ occurring in (2.1) have been observed. And, variables prior to $y_{i,1-p_0}$ and $x_{i,1-p_m}^{(m)}$ and after $y_{i,T}$ and $x_{i,T}^{(m)}$ for $m = 1, ..., M$ are not available. So, just for the sake of simplicity, we assume the panel data set to be balanced. However, particular forms of unbalancedness can easily be accommodated; in the illustration of Section 5 we will use an unbalanced data set. Unobservable are the $p_0 + \sum_{m=1}^{M}(p_m + 1)$ slope coefficients, the $T-1$ time-effects τ_t, the N individual effects η_i and the NT disturbances $\varepsilon_{i,t}$. The model does not include an overall intercept, as the individual effects establish unrestricted individual specific intercepts. The restriction $\tau_1 = 0$ has been imposed to avoid dummy-trap identification problems. Hence, $\tau_2, ..., \tau_T$ represent time-effects in deviation from the individual intercepts. T is supposed to be relatively small, often a one digit number. Therefore the time-effects τ_t can be treated as coefficients of the fixed dummy variables $d_{i,t}^{(s)}$ and estimated jointly with all slope coefficients. As we assume N to be relatively large and approaching infinity in an asymptotic analysis the individual effects will be treated differently.

2.2.1. Classification of variables

A researcher does not only have to make a decision on which explanatories $x^{(m)}$ should be included in the model, but also whether the observed phenomena, including y, should first be transformed, for instance by taking logs or not, measuring them in constant prices or not, scaling them (by size of the population or size of the firm) etc. In this way the linearity in its coefficients does not preclude model (2.1) to represent a relationship which is nonlinear in its basic underlying explanatory phenomena.

Self-evidently, the regressand $y_{i,t}$ is endogenous, because it is determined within the model, implying that it is directly affected by the unexplained component, the random disturbance $\varepsilon_{i,t}$. The regressors of a model can be classified as being either endogenous, predetermined or strictly exogenous with respect to $\varepsilon_{i,t}$. We use a classification of the variables $y_{i,t-l-1}$ and $x_{i,t-l}^{(m)}$ for $l \geq 0$, which presupposes that: (i) the M explanatories and the lag-orders $p_0, ..., p_M$ have been chosen such that the $\varepsilon_{i,t}$ are serially uncorrelated, i.e. $E(\varepsilon_{i,t}\varepsilon_{i,s}) = 0 \; \forall t \neq s$, and (ii) the current unobserved disturbances $\varepsilon_{i,t}$ are in fact innovations with respect to all variables of the model observed prior to t. This implies $E(x_{i,t-l}^{(m)}\varepsilon_{i,t}) = 0$ and $E(y_{i,t-l}\varepsilon_{i,t}) = 0$ for $l > 0$. Then the classification of the variables with respect to $\varepsilon_{i,t}$ is defined as follows:

- self-evidently, all variables $y_{i,t-l}$ for $l > 0$ are predetermined;
- a variable $x_{i,s}^{(m)}$ is classified as strictly exogenous if $E(x_{i,s}^{(m)}\varepsilon_{i,t}) = 0 \; \forall t, s$;
- a variable $x_{i,t}^{(m)}$ is predetermined if $E(x_{i,s}^{(m)}\varepsilon_{i,t}) = 0$ for $s \leq t$ and possibly nonzero otherwise;
- a variable $x_{i,t}^{(m)}$ is endogenous if $E(x_{i,t}^{(m)}\varepsilon_{i,t}) \neq 0$.

Note that if $x_{i,t}^{(m)}$ is endogenous then $x_{i,t-l}^{(m)}$ is predetermined for $l > 0$ and if $x_{i,t}^{(m)}$ is predetermined then $x_{i,t-l}^{(m)}$ is predetermined too for $l > 0$.

Endogeneity of a regressor $x_{i,t}^{(m)}$, say $x_{i,t}^{(1)}$, may occur for various reasons. Assume that an appropriate model for $x_{i,t}^{(1)}$ is

$$x_{i,t}^{(1)} = \textstyle\sum_{l=1}^{p_1^*}\gamma_l^* x_{i,t-l}^{(1)} + \sum_{l=0}^{p_0^*}\beta_l^* y_{i,t-l} + ... + \sum_{s=2}^{T}\tau_s^* d_{i,t}^{(s)} + \eta_i^* + \varepsilon_{i,t}^*. \qquad (2.2)$$

If either $\beta_0^* \neq 0$ or $E(\varepsilon_{i,t}\varepsilon_{i,t}^*) \neq 0$ then $E(x_{i,t}^{(1)}\varepsilon_{i,t}) \neq 0$ and $x_{i,t}^{(1)}$ is endogenous with respect to $\varepsilon_{i,t}$. When $\beta_0 \neq 0$ and $\beta_0^* \neq 0$ then $y_{i,t}$ and $x_{i,t}^{(1)}$ are jointly determined by simultaneous relationships expressing two-way causality. In case model (2.1) suffers from omitted time-varying regressors, or when $x_{i,t}^{(1)}$ is affected by measurement errors, this may also lead to $E(x_{i,t}^{(1)}\varepsilon_{i,t}) \neq 0$ and, more seriously, may also undermine the assumed innovation assumption regarding $\varepsilon_{i,t}$. If $\beta_0^* = 0$ and $E(\varepsilon_{i,t}\varepsilon_{i,t}^*) = 0$, whereas at least one coefficient $\beta_l^* \neq 0$ (for $1 \leq l \leq p_0^*$) then $x_{i,t}^{(1)}$ is predetermined with respect to $\varepsilon_{i,t}$. In that case $x_{i,t}^{(1)}$ is affected by lagged feedback from $y_{i,t}$, whereas the feedback from $y_{i,t}$ (or $\varepsilon_{i,t}$) is instantaneous when $x_{i,t}^{(1)}$ is endogenous. Similar feedbacks may also occur more indirectly, when some of the further explanatory variables of $x_{i,t}^{(1)}$ are affected (probably via $y_{i,t-l}$) by $\varepsilon_{i,t-l}$ for some $l \geq 0$. Direct and indirect as well as instantaneous and lagged feedbacks seem very relevant for actual economic behavioral relationships, so the modelling strategy to be developed should be able to cope with them. However, at the same time this strategy aims to model just the essentials of the autonomous single structural relationship for $y_{i,t}$, leaving open many aspects of relationships such as (2.2) and the complete system in which the single relationship for $y_{i,t}$ operates.

Situations are rare where (dynamic) economic theory is so explicit that it straightforwardly implies the actual true classification of regressors $x_{i,t}^{(m)}$ with respect to $\varepsilon_{i,t}$. Therefore, as long as no convincing theory or empirical evidence points into a different direction, the safe choice is to assume each regressor $x_{i,t}^{(m)}$ to be endogenous, because that requires abstaining from moment conditions which would only be valid in case $x_{i,t}^{(m)}$ would be predetermined or exogenous. Since it makes sense to adopt for an endogenous regressor $x_{i,t}^{(m)}$ the moment conditions $E(x_{i,t-l}^{(m)}\varepsilon_{i,t}) = 0$ for $l > 0$ only when the $\varepsilon_{i,t}$ are serially uncorrelated innovations indeed, it is of paramount importance when specifying (2.1) to effectuate: (a) avoiding the omission of relevant regressors $x_{i,t}^{(m)}$; (b) including a sufficient number of lagged regressors $x_{i,t-l}^{(m)}$ and $y_{i,t-l}$ $(l > 0)$; and (c) choosing an appropriate transformation for $y_{i,t-l}$ and all $x_{i,t-l}^{(m)}$ $(l \geq 0)$ to avoid specifying a deficient functional form.

That pure autoregressive forms of serial correlation can be avoided by choosing appropriately high lag orders for all regressor variables can be shown easily as follows. Suppose that in model (2.1) we actually have that $\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + \xi_{i,t}$ where $\xi_{i,t}$ is serially uncorrelated and $\rho \neq 0$. Hence, $E(\varepsilon_{i,t}\varepsilon_{i,t-1}) \neq 0$ and therefore $E(y_{i,t-1}\varepsilon_{i,t}) \neq 0$, so $y_{i,t-1}$ is actually endogenous now with respect to $\varepsilon_{i,t}$ and the latter is not an innovation. Likewise, if $x_{i,t}^{(m)}$ is endogenous then under serial correlation $E(x_{i,t-1}^{(m)}\varepsilon_{i,t}) \neq 0$ too, so $\varepsilon_{i,t}$ is no longer an innovation with respect to $x_{i,t-1}^{(m)}$ either. In this situation of simple first-order autoregressive disturbances it can be shown easily (see section 3.3) that by increasing all orders $p_0, ..., p_M$ by 1 a model is obtained where the disturbance is no longer given by $\varepsilon_{i,t}$ but by $\xi_{i,t}$, so that in that reparametrized model $y_{i,t-1}$ and $x_{i,t-1}^{(m)}$ are predetermined again. In case of higher-order forms of serial correlation the lag orders should be increased by more than one. This trick will only be effective if $\xi_{i,t}$ is an innovation with respect to $y_{i,t-l}$ and $x_{i,t-l}^{(m)}$ for $l > 0$, which could be at odds with more serious problems than pure autoregressive forms of serial correlation, like omitted regressors, measurement errors, wrong functional form and using proxy variables.

2.2.2. Implementation of GMM

Applying GMM directly to model (2.1) requires instrumental variables which are uncorrelated with $\eta_i + \varepsilon_{i,t}$. Since (lags of) $y_{i,t}$ are, and those of $x_{i,t}^{(m)}$ will, usually be correlated with η_i, the most common implementation of GMM using internal instruments involves estimating the model in first differences. This transformation removes the unobserved η_i from the equation (but not necessarily from its regressors, see stage 10 in section 4.2). Denoting $y_{i,t} - y_{i,t-1} = \Delta y_{i,t}$ etc., it yields

$$\Delta y_{i,t} = \textstyle\sum_{l=1}^{p_0} \gamma_l \Delta y_{i,t-l} + \sum_{m=1}^{M} \sum_{l=0}^{p_m} \beta_l^{(m)} \Delta x_{i,t-l}^{(m)} + \sum_{s=2}^{T} \tau_s \Delta d_{i,t}^{(s)} + \Delta\varepsilon_{i,t}, \tag{2.3}$$

where $i = 1, ..., N$ and $t = 2, ..., T$. Given the adopted classification of the variables moment conditions for estimating (2.3) with disturbance $\Delta\varepsilon_{i,t}$ can be obtained from particular lags and leads of the internal variables of (2.1), namely:

$$\left.\begin{array}{l} E(y_{i,s}\Delta\varepsilon_{i,t}) = 0 \text{ for } s \leq t-2, \\ E(x_{i,s}^{(m)}\Delta\varepsilon_{i,t}) = 0 \text{ for } s \leq t-2 \text{ if } x_{i,t}^{(m)} \text{ is endogenous with respect to } \varepsilon_{i,t}, \\ E(x_{i,s}^{(m)}\Delta\varepsilon_{i,t}) = 0 \text{ for } s \leq t-1 \text{ if } x_{i,t}^{(m)} \text{ is predetermined with respect to } \varepsilon_{i,t}, \\ E(x_{i,s}^{(m)}\Delta\varepsilon_{i,t}) = 0 \text{ for } \forall s \text{ if } x_{i,t}^{(m)} \text{ is exogenous with respect to } \varepsilon_{i,t}. \end{array}\right\} \tag{2.4}$$

Note that, to avoid confusion, we stick to the adopted classification of regressors which just refers to the underlying substantive behavioral model with disturbances $\varepsilon_{i,t}$.[1]

Because we assumed below (2.1) that the dependent variable has been observed over the range $t = -p_0 + 1, ..., T$, the first line of (2.4) seems to imply that just p_0 instrumental variables, namely $y_{i,t-2}, ..., y_{i,t-1-p_0}$ $(\forall i, t = 2, ..., T)$ can be constructed from the dependent variable. However, due to the panel structure of the data and the chosen asymptotic sequence, where T is finite and $N \to \infty$, these p_0 variables can be unraveled to the instrumental variables $d_{i,t}^{(s)} y_{i,l}$, for $s = 2, ..., T$ and $l = 1 - p_0, ..., s-2$ (again $\forall i, t = 2, ..., T$). Hence, from lags of $y_{i,t}$ we can in fact extract $\Sigma_{s=2}^{T}(s - 2 + p_0) = (T-1)(T/2 + p_0 - 1)$ moment conditions. These are valid because applying the law of large numbers yields $\text{plim}_{N\to\infty} N^{-1}T^{-1}\Sigma_{i=1}^{N}\Sigma_{t=2}^{T} d_{i,t}^{(s)} y_{i,l}\Delta\varepsilon_{i,t} = T^{-1}\,\text{plim}_{N\to\infty} N^{-1}\Sigma_{i=1}^{N} y_{i,l}\Delta\varepsilon_{i,s} = T^{-1}E(y_{i,l}\Delta\varepsilon_{i,s}) = 0$.

In the same vain it follows from the second line of (2.4) that for an endogenous regressor $x_{i,t}^{(m)}$ one can obtain $\Sigma_{s=2}^{T}(s-2+p_m) = (T-1)(T/2 + p_m - 1)$ unraveled instrumental variables $d_{i,t}^{(s)} x_{i,l}^{(m)}$ from the moment conditions $E(d_{i,t}^{(s)} x_{i,l}^{(m)} \Delta\varepsilon_{i,t}) = 0$, for $l = 1 - p_m, ..., s-2$. If this regressor were predetermined, $T-1$ extra instruments $d_{i,t}^{(t)} x_{i,t-1}^{(m)}$ can be obtained, corresponding to the extra moment conditions $E(d_{i,t}^{(t)} x_{i,t-1}^{(m)} \Delta\varepsilon_{i,t}) = 0$.

[1] To find all valid moment conditions in terms of internal variables for estimating (2.3) the classification of the variables in (2.1) is sufficient and decisive. So, if we indicate a variable as endogenous, predetermined or exogenous by default this will be with respect to the disturbances ε_{it} of the behavioral model. That in the estimated model the regressor $\Delta y_{i,t-l}$ can be classified as predetermined with respect to $\Delta\varepsilon_{it}$ only for $l > 1$, whereas it could be called endogenous for $l = 1$, may be valid, but is confusing at the same time, so is better avoided.

Whereas, from an exogenous regressor many more than $T-1$ extra valid instruments can be generated than from a predetermined one, namely also $E(d_{i,t}^{(s)}x_{i,l}^{(m)}\Delta\varepsilon_{i,t})=0$ for $s=2,...,T$ and $l=s,...,T$. However, from these, the validity of the $T-1$ extra moment conditions $E(d_{i,t}^{(t)}x_{i,t}^{(m)}\Delta\varepsilon_{i,t})=0$ are especially crucial to distinguish an exogenous variable from a predetermined one. Those for $l>s$ will often yield relatively weak instruments, because these future values will usually not establish substantial extra explanatory power additional to that obtained from current and past variables $d_{i,t}^{(s)}x_{i,s}^{(m)}$ for $s\leq t$ in the so-called first-stage regressions for any of the explanatory variables of (2.3). Although according to asymptotic theory any additional valid instrumental variable would improve estimator efficiency, it is well-known that in finite samples using many instruments may in fact aggravate estimator bias. Therefore, and because they are expected to be weak, the instruments $d_{i,t}^{(s)}x_{i,l}^{(m)}$ with $l>s$ obtained from an exogenous regressor may usually better be omitted from the employed set of instruments.

Given the abundance of available unraveled instrumental variables in a panel data context, all given by interactions between time-dummies and current or lagged regressors, it seems a good idea more generally to reduce their number. To that end there are basically two strategies, namely collapsing (see Roodman, 2009b) and curtailing (see, for instance, Bun and Kiviet, 2006). The unraveled instruments obtained from the lagged dependent variable can, by taking particular linear combinations, be collapsed to p_0-1 plus $T-1$ instrumental variables. The latter are $\Sigma_{s=l}^{T}d_{i,t}^{(s)}y_{i,s-l}$, for $l=2,...,T$, and if $p_0>1$ the former are $y_{i,t-l-1}=\Sigma_{s=2}^{T}d_{i,t}^{(s)}y_{i,s-l-1}$ for $l=2,...,p_0$ and again $t=2,...,T$. The unraveled instruments for $x_{i,t}^{(m)}$ can be collapsed similarly.

An uncomfortable aspect of collapsing seems that in the implied first-stage regression the different time-series observations are confronted now with autoregressive specifications of different orders. When $p_0=1$ then these orders are $t-1$ for observations $t=2,...,T$ respectively. Hence, Δy_{i2} is regressed on just y_{i0}, whereas Δy_{iT} is regressed on $y_{i,T-2},...,y_{i0}$, whilst the one and only coefficient in the regression for $t=2$ is forced to be equivalent to the coefficient of $y_{i,T-2}$ for $t=T$. By curtailing one achieves that these autoregressive specifications have a chosen maximum order, say $1\leq\bar{p}<T-1$, and that apart from limiting the lag order no further coefficient restrictions are being imposed in the first-stage regressions. Then the instrumental variables are $d_{i,t}^{(s)}y_{i,l}$, for $s=2,...,T$ and $l=\max(1-p_0,s-1-\bar{p}),...,s-2$. When $\bar{p}\leq p_0$ their number will be $(T-1)\bar{p}$ and when $\bar{p}>p_0$ then $(\bar{p}-p_0+1)(\bar{p}-p_0)/2$ less. Hence, curtailing tends to use more instruments than collapsing, but seems less strained too, because in the first-stage regression all observations $t\geq\bar{p}-p_0+2$ will be fitted to the same number of $\bar{p}$ regressors, namely $d_{i,t}^{(t)}y_{i,l}$ (for $l=t-\bar{p}-1,...,t-2$), without coefficient restrictions. The two principles can also be combined, see Kiviet et al. (2017, p.10) for more details. Of course, the unraveled instruments obtained from $x_{i,t}^{(m)}$ can be curtailed and/or collapsed similarly.

As follows directly from true relationship (2.3), for the instrumentation of $\Delta y_{i,t-l}$ for $l\geq1$, the inclusion in the set of instruments of (unraveled versions of) $x_{i,t-1}^{(m)}$ and $x_{i,t-2}^{(m)}$, next to $y_{i,t-l}$ for $l\geq2$, seems crucial. However, practitioners seldomly stick to this. In classic static nonpanel IV models it makes sense to instrument exogenous regressors just by themselves, whereas the instrumentation of endogenous variables requires additional external instrumental variables. This has wrongly induced the habit in dynamic panel model estimation to instrument exogenous regressors just by themselves too, implying using just the $p_m + 1$ orthogonality conditions $E(\Delta x^{(m)}_{i,t-l}\Delta\varepsilon_{i,t}) = 0$ for $l = 0, ..., p_m$. In case $p_m = 0$ this simplifies to using just the instrument $\Delta x^{(m)}_{i,t}$, whereas the very specification of model (2.3) highlights that to instrument regressor $\Delta y_{i,t-1}$, in addition to $y_{i,t-l}$ for $1 < l \leq p_0 + 2$, also variables $x^{(m)}_{i,t-l}$ for $1 < l < p_m + 2$ will be most useful.

From the above it follows that, due to the panel structure of our data, finding a huge number of instruments seems relatively easy. Therefore, practitioners are usually not forced to seek external instruments, which are variables uncorrelated with the current disturbance in the estimation equation which have not been obtained by lagging or transforming the regressors of the model. However, although relatively uncommon in practice, nothing opposes the use of external instruments when analyzing panel data, provided they are valid and relevant. This requires that they have nonzero coefficients in equations like (2.2).

2.3. Adopting an initial model specification and set of moment conditions

For a great number of reasons the search for an initial acceptable general model is quite difficult. A major handicap is that economic theory is usually not very specific regarding all variables that are relevant for the causal explanation of $y_{i,t}$, nor regarding the proper functional form of the relationship, and certainly not with respect to the characteristics of the dynamics. So, it is usually the task of the applied econometrician to asses most of these issues empirically. Though, an additional difficulty is that econometric tests on the validity of instruments and on the adequacy of a chosen specification of the model may commit type I or type II errors (rejecting a true hypothesis or not rejecting a false hypothesis) with possibly substantial probabilities and they may therefore frequently misguide.

The motivation for promoting in the present context to include in the initial model specification relatively many (lagged) regressors while using very few unlagged and first-order lagged regressor variables as instruments in the first-differenced model, is based on trying to avoid inconsistency of estimators.[2] The stimulus for this policy stems from the following characteristics of the statistical tools to be used. Econometric tests are available to examine the (in)validity of a subset of instruments when added to an already accepted set of instruments, and also to examine whether a subset of regressors should (not) be excluded from an already accepted model specification. However, for a proper interpretation these tests require that the already accepted model specification is genuinely adequate and that the already accepted set of instruments is really valid for this accepted model specification. These two test tools lead to generally uninterpretable results when applied in a situation where the already accepted model and accepted instrument set produce inconsistent estimates. Therefore, it is crucial to find an initial specification of the model and set of instruments on which evidence can be produced which provides trust in their joint adequacy. Part of this is avoiding excluding any possibly relevant explanatories and avoiding at the same time exploiting any possibly invalid instruments. As an unavoidable consequence, in order to achieve consistency, one has to sacrifice some efficiency initially by unknowingly including regressors which may actually be redundant, and by yet abstaining from using particular instruments which

[2] See also Hendry (1995), who pleads for a general-to-specific methodology in the context of econometric time-series analysis.

may actually be valid and effective.

This quite general initial model specification with relatively restricted set of instruments, for which no evidence should have been found regarding invalidity of some of its instruments, nor regarding possibly omitted regressors (including those that may represent a wrongly chosen functional form, and those representing non constant coefficients due to omitted interaction terms), nor regarding detrimental serial correlation of its errors (due to wrongly omitting lagged or other regressors), we will address as the general maintained statistical model (MSM). For its results one should neglect for the moment that many of its coefficients may not appear as significant due to the relative inefficiency of its estimates.

In the present context we have found an acceptable MSM when a possibly not very parsimoniously specified dynamic regression model, in which most of its current (non lagged) regressors are treated as endogenous (i.e. possibly affected by immediate feedback from the dependent variable), after being estimated in first-differenced form (to get rid of individual effects) by Arellano-Bond GMM estimation (either 1-step with heteroskedasticity robust standard errors, or 2-step with Windmeijer-corrected standard errors), yields residuals which do show first-order serial correlation (due to the first differencing) but no significant second order serial correlation, and for which the test of the over-identification restrictions, addressed as the Hansen test in Xtabond2[3], has a reasonably large p-value (what seems large in this context will be discussed in subsection 2.4). Employing at this stage GMM system estimation according to Blundell and Bond (1998) would be unwise in general, because this would assume validity of some additional instruments in the level equation, whereas the tool to test their validity presupposes validity of the Arellano-Bond instruments, which are yet still under scrutiny. So, first all emphasis should be on verification of the latter. Blundell-Bond estimation and verification of its possible validity should be postponed until a later stage in the modelling process.[4]

Hence, initially, GMM estimates and tests for the first-differenced model should be used until a satisfactory MSM has been obtained, and next it could be examined whether possibly some further effective internal instruments (obtained from first-lagged or current regressors) can safely be adopted as well, while some weak ones may successfully be removed (possibly by collapsing or removing long lags), together with removing from the model specification some ostensibly redundant regressors (when they are found to produce very small t-ratios or correspondingly high p-values) or to impose other coefficient restrictions when strongly supported by the data.

2.4. The major dilemmas faced

In trying to avoid as much as possible to impose restrictions on the model that have not yet been tested (and thus have not yet been rejected nor confirmed) one faces the following dilemma. When using GMM techniques to a model with some endogenous regressors it is unavoidable to impose some untestable exclusion restrictions on the model

[3] Note that the test addressed as the Sargan test in Xtabond2 presupposes (conditional) homoskedasticity of the disturbances, so this can only be interpreted when no evidence of heteroskedasticity has been found.

[4] As long as Arellano-Bond results are unsatisfactory, applying Blundell-Bond does not make sense. Therefore it is unfortunate that Blundell-Bond is the default technique in Xtabond2.

in order to satisfy the order condition of identification. The reason is that we need at least as many instrumental variables as we have regressors in the model, whereas an endogenous regressor as such cannot be used as an instrument, because it will be correlated with the error term due to instantaneous feedback. In fact, when estimating the model in first-differenced form, also the first lag of an endogenous variable cannot be used as instrument, and neither can the current level of predetermined regressors. Therefore, particular lagged variables to be used as instruments should not occur as regressor at the same time, because otherwise the number of instruments obtained from lagged variables would not outnumber the regressors. Note that to formally test validity of their exclusion from the set of regressors, one should first include them, but that extended model can no longer be estimated (due to having fewer instruments now than regressors), so these unavoidable exclusion restrictions do necessarily imply untestable restrictions on the dynamic specification of the model. They would only be testable when valid external instruments (not obtained by transformation of included regressor variables) would be available, but then the validity of these external instruments becomes untestable, because if one includes them for testing whether they are uncorrelated with the model errors the resulting extended model is no longer identified (its number of regressors exceeds the number of instruments).

There is a strong interconnection between tests for the validity of instrumental variables not included in the regression and tests about whether these instrumental variables have been wrongly omitted from the regression, see Kiviet (2017). This induces why we can test only the over-identification restrictions, while implicitly assuming that the untested just-identifying restrictions (equal in number to the number of unknown coefficients of the model) are all simply valid.[5] The above should make clear that one always has to keep an open mind regarding the possibility that an accepted maintained statistical model (MSM) may actually be false, because many of its underlying exclusion restrictions and moment conditions have not and cannot be tested (without adopting further untestable conditions) and thus have to be accepted simply (but possibly wrongly) in good faith.

At first sight the concern just mentioned seems less worrying for panel data analysis than it is for a regression analysis based purely on either cross-section data or on time-series, because the extra dimension of microeconometric panel data enables to extract from one regressor a multitude of instruments which suggests that over-identification is ubiquitous.[6] Including all these instruments (interactions of time-dummies with current and lagged variables) as regressors would generalize the specification such that each explanatory variable for all its lags would have for each particular time period its own unique coefficient. In fact, this would imply incorporating into the model the interactions of all time dummies with all available lagged regressors, which then would undo the over-identification. In practice, most if not all of these interaction terms are implicitly assumed to have equal coefficients. In the presence of endogenous regressors, adopting restrictions on some of those interactions is simply unavoidable, and it is untestable without adopting other untestable assumptions. So, also in a panel context adopting

[5] For an escape route to test just-identifying restrictions in nonpanel models, see Kiviet (2019).

[6] Below (2.4) it is shown that when $p_m = 1$ and regressor $x_{i,t}^{(m)}$ is endogenous it yields $(T-1)T/2$ instruments, and $T-1$ more when it is predetermined, and still more when it is exogenous. This is not yet very impressive when $T = 3$ (which is the minimum required). Then each regressor produces at least 3 instruments, but at least 6 when $T = 4$, and 10 when $T = 5$, etc.

(explicitly or implicitly) untestable exclusion restrictions in order to achieve identification is simply unavoidable. In fact, a similar situation occurs for pure cross-section and time-series data. Here too one can proliferate the number of instruments, for instance by taking their square, higher-order powers or a range of other nonlinear transformations or their interactions with dummy variables. This only yields (most probably weak) over-identification if one is willing to make the untestable assumption that at the same time sufficient of these extra instrumental variables have zero coefficients in the regression.

Hence, in practice one will always be faced with making a choice between generality of the specification and imposing some untested restrictions. This concerns the inclusion or not of particular variables, choosing their maximum lag order and the occurrence of (lagged) variables in general interaction terms as well (the latter are addressed in subsection 3.4). In designing the MSM one has to trade off the risks of devastating time-varying unobserved heterogeneity, due to omitted regressors, and of inefficient inference, due to redundant regressors and unexploited valid over-identifying restrictions. Because one cannot test all implicit restrictions imposed on a candidate MSM one has to keep in mind that, although subjective elements in its design can to some degree be mitigated, they are in fact unavoidable.

On top of all the dilemmas just mentioned there is yet another one. In contrast with the foregoing, which pleaded to include many potential explanatories, this one urges for not including too many potential candidate regressors for reasons which again have to do with identification of the single structural relationship that we want to establish. In addition to the necessary order condition (having at least as many instruments as unknown coefficients) we have to satisfy the necessary and sufficient rank condition for identification, and this is certainly not as easily satisfied as the necessary order condition. The single structural equation that we want to analyze is one from a system which specifies for each endogenous variable in that system their single structural equation. Although the single structural equation that we want to estimate is nested in a comprehensive single equation model, which includes all endogenous, predetermined and exogenous variables that occur in the full system, such a heavily overspecified equation is certainly not identified according to the rank condition, because all individual structural equations of that system are nested in the very same comprehensive model. Without imposing relevant coefficient restrictions it cannot be distinguished from the other structural equations. In a panel data context it may seem possible to estimate the comprehensive model because sufficient lagged variables interacted with time dummies may be available to meet the necessary but insufficient order condition. However, interpretation of its estimates with respect to the structural equation for the single dependent variable that we want to model is not possible as long as the sufficient rank condition has not been fulfilled, which requires imposing a sufficient number of the appropriate coefficient restrictions which make the equation of our interest unique amongst the other equations of the system.

We should highlight here a related hazard. Consider again model (2.1) and assume $\beta_0^{(1)} \neq 0$. Then any specification for a model for $y_{i,t}$ which omits $x_{i,t}^{(1)}$ as a regressor is misspecified as a representation of the structural model, but it is not necessarily misspecified as such. When model (2.2) is substituted in (2.1) then $x_{i,t}^{(1)}$ is eliminated and it brings in all other variables in the structural equation for $x_{i,t}^{(1)}$ not yet occurring in the structural equation for $y_{i,t}$. The resulting equation is something half-way the structural form equation and the reduced form equation for $y_{i,t}$. In the latter all endogenous regressors have been eliminated, and only predetermined and exogenous regressors remain. To make sure that our specification search yields in the end the structural equation for $y_{i,t}$ we should avoid including in its initial deliberately overspecified form any regressors which occur in other structural equations of the same system but have zero coefficients in the structural equation for $y_{i,t}$. Otherwise we face the risk that, even when the adequate specification of the structural equation of $y_{i,t}$ is nested in our MSM, when aiming for parsimony, we may remove endogenous explanatories and leave in indirect determinants of $y_{i,t}$. The latter have an effect on it only via the endogenous explanatories, whereas these endogenous regressors apparently belong to the direct explanatories. It are the direct explanatories which constitute the autonomous structural relationship for $y_{i,t}$.

So, although it seems a good idea to opt for a methodology in which one starts off with a relatively general model specification, both regarding the dynamics and the inclusion of further explanatory variables and allowing their endogeneity by abstaining from using particular instruments, at the same time imposing a substantial number of coefficient restrictions by omitting particular variables as regressors is unavoidable too. Not only in order to have a sufficient number of instruments. It is also required in order to keep focus on establishing the structural causal relationship for $y_{i,t}$. Hence, many nested and nonnested specifications of models for $y_{i,t}$ may each qualify as an appropriate candidate MSM, because they are individually statistically acceptable; however, not all of them may contain the structural model for $y_{i,t}$ as a special case. Moreover, the fewer restrictions one imposes on the candidate MSM and the fewer instrumental variables one exploits, the larger the coefficient standard errors will be. This will mitigate the power of tests on the significance of coefficients but also on the validity of instruments, whereas the information to be exploited in a successful specification search, as to be developed below, has to come from the correct (non-)rejection of series of such tests. So, these individual tests having substantial test power (low type II error probability), whilst maintaining a well controlled type I error probability, is crucial.

2.5. P-values as beacons: opportunities and shortcomings

In the specification and selection search strategy to be designed in Section 4 the sequential application of various particular hypotheses test procedures will yield p-values which will constitute the key factors on which next steps in the search process will be based. All these separate steps will be associated with testing the (in)significance of a particular subset of parameters, where the outcome will be interpreted as either including or excluding the variables associated to these parameters to/from either the regression or the set of instruments. This (in)significance, indicating the (in)compatibility of the observed data with all these tested parameters having value zero[7] or at least one of them having a nonzero value, is expressed in the p-value, a scalar between zero and one. Its magnitude will determine whether the tested hypothesis will be rejected or not. In the latter case the strategy yet to be developed will imply taking the decision to impose the parametric restrictions tested, i.e. accepting the tested null hypothesis. Although in this way the resulting chain of decisions is in serious conflict with the classic statistical theory

[7] Testing whether each parameter of a subset of parameters has a particular real value can easily be converted into a test in which for a similar subset it is tested whether all its individual elements are zero or not.

of hypothesis testing it can nevertheless be defended using simple pragmatic arguments. These are: (a) there is no straight-forward alternative practical strategy; (b) as shall be indicated in the next subsection sub-strings of the chain of decisions to be made can be embedded in a cautious theory of testing sequentially superimposed or juxtaposed hypotheses; and (c) by being fully transparent about how all individual decisions in the structured sequence came about, its results will allow constructive criticism and at the same time candidly reveal both its fragility and purport. By being completely open about its subjective and possibly impertinent aspects other researchers are facilitated to uncover all its incongruities and come up with alternative proceedings and explanations, which in their turn will undoubtedly be debatable too.

Given the purpose for which we will use them p-values of test statistics have two major shortcomings. Firstly, they do not express –nor can they be converted into– an index accurately indicating how likely it is that either the tested null hypothesis (particular parameters having value zero) or its denial (some of those parameters being nonzero) is true. Secondly, in the present context of dynamic panels the calculated p-values involve asymptotic approximations, because the probability distribution by which they have been calculated is the relevant probability distribution for the test statistic concerned only if the sample were infinitely large in the cross-section dimension. We will now discuss both these shortcomings and some of their consequences one by one.

In a recent editorial statement of the American Statistical Association Wasserstein and Lazar (2016) clarify some principles regarding p-values, all the time just considering situations in which the second shortcoming is absent. For panel data modelling of behavioral relationships this would require the unrealistic situation that all regressors are strictly exogenous (so no lagged dependent variable regressors nor any further regressors affected by feedbacks should occur in the relationship under study), whereas the distribution function of the disturbances should be fully known, apart from a scale factor. Under these very restrictive circumstances the p-value of a test statistic is the exact probability under a specified statistical model fully obeying the restrictions specified by the null hypothesis that it is equal to or more extreme than its observed value. Hence, the lower such a p-value is the more incompatible the observed data seem with the specified statistical model under the null hypothesis. Then, apparently, the data may stem from a different statistical model. This could either be the specified statistical model under the alternative hypothesis, where the latter just allows the tested subset of parameters to have nonzero values, or a completely different statistical model. Only when truth of the adopted statistical model is maintained (meaning: its truth is beyond dispute) then rejection of the null hypothesis directly implies acceptance of the alternative hypothesis (being: not all tested parameters have value zero).

Thus, when a MSM has not yet been assessed, a low p-value as such does not necessarily carry much information on how an adequately specified model should look like. Although a high p-value indicates substantial compatibility of the specified statistical model under the null with the actual data, it could well be that a rather differently specified statistical model is in fact much more compatible with the data. Even a p-value of 1 does not imply truth of the null hypothesis; it just means that the estimated values of the parameters correspond to their hypothesized values, but this does not imply that their true values correspond to the hypothesized values as well. Also a p-value of zero does not imply truth of the narrowly interpreted alternative hypothesis; it just indicates that the model under the null is badly specified.[8] So, choosing some fixed threshold value between zero and one, which separates the decision to either impose or not the zero values for the tested parameters, seems doomed to fail.

The plead to start off the search strategy by a rather uncontroversial initial model specification with a prudent selection of instruments tries to put the sequence of tests to be performed within a context where validity of the assumptions constituting this initial statistical model can rather safely be adopted. If no obvious shortcomings of this statistical model can be found by testing it against even less restrictive alternatives, then we may adopt it as our general MSM (maintained statistical model). Then the whole further analysis, which aims to increase efficiency, can be interpreted as being conditional on these maintained assumptions. Then rejecting a null hypothesis implies to accept that the parameters tested are not all zero, which constitutes the alternative hypothesis under which estimates are still consistent under the general MSM. The decision to impose either non rejected coefficient restrictions or non rejected orthogonality conditions leads to tightening of the initially adopted MSM. If these imposed restrains are valid indeed then the estimates will still be consistent and also more efficient, but when invalid consistency will be lost in general which may lead to seriously biased estimates.

These two possible consequences (either gaining efficiency, or loosing consistency) should determine which threshold value for the p-value should be used when deciding to reject a null hypothesis or to accept it. Choosing for this the habitual significance level $\alpha = 0.05$ would now in most cases be weird as we shall argue. For the moment we will disregard the complex consequences of both pretesting (the fact that a single test may have been preceded by decisions taken on the basis of other tests calculated from the very same sample data) and possible inaccuracies due to working with asymptotic p-values instead of exact p-values. Just focusing on one separate test, taking 0.05 as the borderline between reject and accept would limit the probability to incorrectly reject the null hypothesis (commit a type I error) to 5%, with the consequences of a type I error just being to miss out some potential efficiency gains. Whereas not rejecting a false null (commit a type II error) will have the much more serious consequence to accept this false null, leading to inconsistent estimates and misguided inferences. So, avoiding type II errors seems here much more crucial than avoiding type I errors. Hence, it seems of much more importance to limit the type II error probability to some low threshold instead of a type I error. However, unlike the type I error probability, the type II error probability is unknowable in practice, because it depends on the actual yet to be modelled data generating process and its true parameter values, and not on a hypothesized data generating process specified by the tested null hypothesis, as is the case for the type I error probability. When we would employ the extreme threshold $\alpha = 1$ we would reject any null hypothesis and just stick to the initial MSM and exclude to use any further information extracted from the sample regarding the possible absence of contemporaneous or lagged feedbacks and redundancy of particular included regressors. At the same time it would exclude the risk of introducing inconsistency. So, what threshold value seems reasonable?

To find an answer to that question let us get back to the phase of designing the initial specification which should be so general that it seems uncontroversial, but not so general that it establishes an unidentified parametrization of the structural equation

[8] A long list of popular misinterpretations of p-values is provided in Greenland et al. (2016).

that we intend to model. Of course, we should always be aware that it may be the case that the available data set simply lacks particular (proxies for the) variables that are indispensable for modelling the structural equation. So there is always the possibility that an acceptable and valid MSM, which contains the true structural data generating process as a special case, cannot and should not be found.

However, let us suppose for the moment that we have carefully designed an initial candidate MSM and that the p-value for the test of its over-identification restrictions is found to be, say, 0.2. Would that be satisfactory? There is no fully comforting simple objective answer to that question possible. Of course, at first sight we should be much happier with a value of, say, 0.8 or even larger, although even that would not guarantee that we are on the right track. Also, as can easily be learned by running some rather haphazard computations on the basis of a rather arbitrary data set, a value of 0.8 or higher can relatively easily be obtained when the sample size is not very large and when using a huge number of candidate instruments. This will be due to the test having very moderate power (probability to reject a false null) in such circumstances, in combination with the second complication (the asymptotic approximative nature of p-values which we still have to address). Though, on the other hand, supposing that the tested null is really true, so when drawing from the actual null distribution of a test statistic, and then obtaining a value with a tail probability of 0.2 seems not an extremely unlikely event; it is more likely than throwing a six with a die. So, in general, finding for the standard Hansen statistic a value of 0.2 should not automatically lead to discarding the candidate initial model. On that score, though, the same can still be said for a value of 0.15. Hence, an absolute fixed threshold cannot be given, but a test value just slightly above 0.05 does certainly not provide strong evidence in favor of validity of the instruments. Moreover, whether we should be satisfied when a candidate MSM produces a value of, say, 0.15 for the Hansen test also depends on the perceived power of the test. If the power seems very moderate a threshold of 0.15 seems rather small, because it would involve a substantial risk to accept inconsistent estimates.

To elaborate on this, imagine we have the choice between two different test procedures for the same null hypothesis, where one is more powerful than the other. The more powerful one is supposed to produce a more extreme value for its test statistic (in relation to its null distribution) when the null is false, and thus yields a smaller p-value more easily. Using both against the same threshold value will lead less frequently to incorrectly accepting the null by the more powerful test. Hence, the threshold p-value for the more powerful test could be chosen smaller to realize similar risks for both tests, or the less powerful one should be used at a relatively large p-value threshold.

So, arguing what threshold would be reasonable becomes slightly easier in situations where one may hope that a test has relatively good (or bad) power properties. For instance, in a situation where one is testing a large number of over-identification restrictions whereas many of these may actually be valid whereas just a few may be false, it seems self-evident that a test procedure which just tests the validity of the doubtful restrictions while building on the validity of the other valid over-identification restrictions is more powerful than the overall test which tests all restrictions jointly. From this we conclude that for difference in Hansen tests one may in principle use a threshold value closer to (though usually still substantially larger than) 5% than for an overall Hansen test. For the latter we conclude at this stage that a threshold of 30% or even 50% might certainly be worth considering. However, more has to be said about this.

First, we will discuss the consequences of the asymptotic nature of p-values in the present dynamic panel model context (see Appendix A for more details). In the presence of feedback variables, and when the distribution function of the disturbances is unknown and is probably also characterized by an unknown form of heteroskedasticity the null distribution of over-identification and coefficient restriction tests is unknown too, although in large samples it converges to a χ^2 distribution (or to standard normal z when single coefficient restrictions are being tested and one-sided alternatives can be considered). So, when calculating the p-value, the observed test statistic is not confronted with the true relevant null distribution, but with its χ^2 (or z) asymptotic approximation, or after degrees of freedom corrections probably with their F (or t) approximation. Focussing on using the χ^2 or F variants of the tests, in case the actual null distribution is located more to the right (left), then the calculated p-value is systematically too small (large). Hence, in order to realize an intended type I error probability of, say, 0.25 one should use a larger (smaller) threshold than 0.25. Some insights in the actual location of the relevant null distributions in dynamic panel analysis has been obtained from Monte Carlo simulations (see, for instance, the second half of Kiviet et al. 2017). Its discrepancies from the asymptotic approximation depends on many specific characteristics of the actual data generating process, but generally speaking it seems very often the case that the p-values obtained for overall Hansen tests are too small (hence, tend to reject too easily) and those for coefficient tests are too large (thus, tend to accept too easily).

The above text seems to suggest that p-values obtained from observed test statistics to decide on instrument or regressor selection issues should be used by confronting them with a threshold value which should be chosen such that the risks of loosing consistency or miss out efficiency gains seem in some perceived balance. However, we could also try to base these decisions on more classic model selection criteria. In standard regression the choice between a more and a less restricted model can be based on comparing the adjusted coefficient of determination, or on the AIC or BIC. This approach closely corresponds to using a t or F test for the restrictions and using the critical value 1 as threshold: Reject the restricted model when the test statistic is larger than 1 and accept otherwise. In samples with several hundred observations this closely corresponds to using the χ^2 version of the test and using the number of tested restrictions (the degrees of freedom, also being the expectation of the null-distribution) as the threshold value. Hence, neglecting the skewness of the χ^2 distribution, this closely corresponds to rejection if the test statistic is in the right-hand half of the null-distribution and acceptance for a realization in the left-hand half, which is close to a 50% significance level instead of 5%. More precisely, it can rather easily be checked that this corresponds in fact to using the p-value of the test as criterion and employing a threshold which is close to 0.5 indeed when over 100 restrictions are being tested, whereas it is 0.68 when just 1 restriction is tested. A few more of these threshold values are: 0.63 (for 2 restrictions), 0.61 (for 3), 0.58 (5), 0.56 (10), 0.54 (20), 0.53 (50) and 0.52 (for 100). Hence, model selection criteria applied to nested alternatives implicitly use significance levels which are much and much larger than 5%.

Simply assuming that similar thresholds could also be employed in a GMM context, but also taking into account the systematic asymptotic approximation errors and our desire to mitigate type I but especially type II errors, we feel an undoubtedly subjective and yet poorly evidence-based preference to recommend the following thresholds. For the test on the first-order serial correlation coefficient of the disturbances a p-value threshold in the range 0.01-0.05 does seem reasonable, because the test should have substantial power when the actual coefficient is -0.5 instead of zero. For the test on the second-order serial correlation coefficient a threshold in the range 0.05-0.15 might be reasonable, because we expect the test to have only modest power and want to avoid type II errors. For overall Hansen tests with a large number of degrees of freedom we suggest to use a p-value threshold in the range 0.10-0.20, and for incremental Hansen tests with less than 10 degrees of freedom in a candidate MSM a threshold in the 0.05-0.15 range, but a range of 0.30-0.50 for deciding whether an initially as endogenous treated regressor seems actually predetermined or whether an as predetermined treated regressor is actually exogenous. For omitting single regressors from models with satisfactory Hansen tests we feel inclined to use as a threshold for the t or z statistic a value in the range 0.5-1.0 (or a p-value threshold in the range 0.4-0.6). For joint tests on the significance of substantial groups of regressors a p-value threshold of about 0.5-0.7 may be reasonable to balance the desire to improve efficiency and avoid inconsistency. Especially samples where N is really small (say, smaller than 150) or reasonably large (over 2500, say) may certainly motivate modifications of these tentative ranges on thresholds.

2.6. Consequences of sequential testing

In the above and the selection strategy developed below it is suggested to use various test procedures again and again at the different stages of the model selection process while using all the time the same one and only sample. Although this is most common amongst practitioners in the social sciences, at the same time it is hard to structure such practices in such a way that they can be justified as a proper and scientifically sound methodology. Here we will just mention some literature that provides some support to aspects of the overall strategy to be laid out in subsection 4.2.

Spanos (2017) argues that the inductive phase in which misspecification-testing is being used to discover a statistically adequate model by repeated re-specification should be seen separate from –and is also indispensable for– a meaningful next deductive phase in which substantive inferences are being produced by tests on the parameters of the examined structural relationship. In Kiviet and Phillips (1986) and Hendry (1995, p.490) it is argued how overall test size of sequences of individual tests can be controlled in general on the basis of the Bonferroni inequality, and especially when the successive maintained hypotheses in these strings of tests are one by one nested so that the individual test statistics are asymptotically independent.

3. Specification of the dynamics

In this section we will spell out how the inclusion of lagged variables in the regression will determine the dynamic reaction patterns by deriving impact, interim and total or long-run multipliers. We start off in the first subsections by focusing on pure time-series data; for a more extensive treatment see Harvey (1990) or Hendry (1995). Next, generalizing for standard panel data models is reasonably straight-forward. We discuss the major differences in interim multipliers for a partial adjustment model and a more general autoregressive distribute lag model by casting both in error-correction form. In standard panel models all multipliers are homogeneous over all subjects. By introducing (lagged) interaction terms heterogeneity can be accommodated at the cost of relatively few extra parameters. This is examined in the fourth subsection and in the fifth the impact of feedbacks on multipliers is discussed. Finally, in the sixth subsection we develop an algorithm to calculate interim multipliers for dynamic models with interactions.

3.1. ADL models

Consider the simple second-order dynamic time-series regression model ($t = 1, ..., T$)

$$y_t = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \beta_0 x_t + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \varepsilon_t,$$

where we assume that ε_t is white-noise, meaning that $\varepsilon_t \sim (0, \sigma_\varepsilon^2)$ and serially uncorrelated. For the moment we will also assume that scalar variable x_t is strictly exogenous, so $E(x_t\varepsilon_s) = 0 \; \forall t, s$. An acronym for this autoregressive distributed lag model specification is $ADL(2, 2)$.

If x_t were (hypothetically) running at constant finite value $\bar{x}$, then $E(y_t)$ would be constant and finite too (provided the values of γ_1 and γ_2 are such that the model is stable, as will be clarified below), and in fact be given by

$$\bar{y} = \frac{\beta_0 + \beta_1 + \beta_2}{1 - \gamma_1 - \gamma_2}\bar{x}.$$

A permanent increase in $\bar{x}$ by 1 unit, would increase $\bar{y}$ by $TM_x^y = (\beta_0 + \beta_1 + \beta_2)/(1 - \gamma_1 - \gamma_2)$. This is called the total or long-run multiplier of y with respect to x. If x and y are in fact the logs of underlying variables X and Y, then TM_x^y is the long-run elasticity of Y with respect to X.

A more sophisticated way to write down dynamic models is by making use of the lag operator L. This operates on the time index of any variable v_t as follows: $L^h v_t = v_{t-h}$ for integer values of h. Hence, $L^0 v_t = v_t$, $Lv_t = v_{t-1}$ etc. The above model can now be written as

$$(1 - \gamma_1 L - \gamma_2 L^2)y_t = (\beta_0 + \beta_1 L + \beta_2 L^2)x_t + \varepsilon_t.$$

More generally, a linear dynamic regression model with M separate exogenous explanatory variables is the $ADL(p_0, p_1, ..., p_M)$ model, given by

$$\gamma(L)y_t = \textstyle\sum_{m=1}^{M} \beta^{(m)}(L)x_t^{(m)} + \varepsilon_t, \tag{3.1}$$

where polynomial $\gamma(L) = 1 - \gamma_1 L - ... - \gamma_{p_0} L^{p_0}$ and the polynomials $\beta^{(m)}(L) = \beta_0^{(m)} + \beta_1^{(m)} L + ... + \beta_{p_m}^{(m)} L^{p_m}$ have for $m = 1, ..., M$ finite nonnegative integer orders $p_0, p_1, ..., p_m$ respectively, and finite real coefficients γ_h, $h = 1, ..., p_0$ and $\beta_h^{(m)}$ for $h = 0, ..., p_m$.

Such a model is stable when polynomial $\gamma(L)$ has all its roots outside the unit circle. There are no requirements regarding the roots of the polynomials $\beta^{(m)}(L)$. This can be understood by the following reasoning. Let z be an arbitrary possibly complex scalar variable. Now $\gamma(z) = 0$ will have p_0 real or complex roots. Denoting these as $\gamma_1^*, ..., \gamma_{p_0}^*$ then $\gamma(z) = \prod_{l=1}^{p_0}(1 - z/\gamma_l^*)$, where all γ_l^* are a function of $\gamma_1, ..., \gamma_{p_0}$. Assuming that for each m all variables x_t^m run at constant and finite values, then $\gamma(L)E(y_t) = c$ with c some finite constant, whereas $E(y_t) = \prod_{j=1}^{p_0}[1 - (1/\gamma_j^*)L]^{-1}c$. Let $\gamma_j^{**} = 1/\gamma_j^*$. Using $Lc = c$ and $(1 - \gamma_j^{**}L)^{-1} = 1 + \gamma_j^{**}L - (\gamma_j^{**}L)^2 + (\gamma_j^{**}L)^3 - ....$, it follows that $E(y_t)$ will be constant and finite only when all γ_j^{**} are within the unit circle, because otherwise $(1 - \gamma_j^{**}L)^{-1}$ does not converge. If $p_0 = 1$ then $\gamma(L) = 1 - \gamma_1 L$, which has real root

$\gamma_1^* = 1/\gamma_1$, whereas for stability $\gamma_1^{**} = \gamma_1$ should be smaller than 1 in absolute value; this is the familiar requirement for stability of the first-order autoregressive model. Such an AR(1) model is even covariance-stationary if $Var(y_0)$ is such that $Var(y_t)$ is constant for $t \geq 0$.

In what follows we will all the time assume that $\gamma(L)$ has all its p_0 roots $\gamma_1^*, ..., \gamma_{p_0}^*$ outside the unit circle. Thus, we restrict ourselves to dynamically stable behavioral relationships. That does not imply that the variables $x_{i,t}^{(m)}$ are not allowed to be non-stationary through time. It just means that the degree of integratedness of $y_{i,t}$ is the same as that of the highest degree amongst the variables $x_{i,t}^{(m)}$. Because T is supposed to be finite the actual degree of integratedness of the variables is of no concern for the inference techniques to be employed.

3.2. Dynamic multipliers

The **total multiplier** of regressor $x_t^{(m)}$ $(m = 1, ..., M)$ given stable model (3.1) with all regressors $x_t^{(m)}$ exogenous is simply $\beta^{(m)}(1)/\gamma(1) = \sum_{h=0}^{p_m} \beta_h^{(m)}/(1 - \sum_{h=1}^{p_0} \gamma_h)$, or, if we define $\delta^{(m)}(L) = \beta^{(m)}(L)/\gamma(L)$, then

$$TM_m = \delta^{(m)}(1) = \textstyle\sum_{d=0}^{\infty} \delta_d^{(m)}. \tag{3.2}$$

It expresses the long-run effect on $E(y_t)$ of a unit change in $\bar{x}^{(m)}$, whereas all other regressors remain constant (ceteris paribus). The total multiplier is zero when $\beta^m(1) = 0$, implying that at least one of the roots of the lag polynomial $\beta^{(m)}(L)$ is unity. Then we can factorize $\beta^m(L) = (1 - L)\beta_*^m(L) = \Delta\beta_*^m(L)$, where the order of $\beta_*^m(L)$ is $p_m - 1$. $TM_m = 0$ occurs for $p_m = 1$ when $\beta_1 = -\beta_0$ and for $p_m = 2$ when $\beta_2 = -(\beta_0 + \beta_1)$. $TM_m = 1$ occurs when $\gamma(1) = \beta^m(1)$ or when $\sum_{h=1}^{p_0} \gamma_h + \sum_{h=0}^{p_m} \beta_h^{(m)} = 1$.

As a rule the rational lag polynomial $\delta^{(m)}(L) = \beta^{(m)}(L)/\gamma(L)$ is of infinite order, where $\delta^{(m)}(L) = \sum_{d=0}^{\infty} \delta_d^{(m)} L^d$, with

$$\left.\begin{aligned} \delta_0^{(m)} &= \beta_0^{(m)}, \\ \delta_d^{(m)} &= \textstyle\sum_{l=1}^{\min(d,p_0)} \gamma_l \delta_{d-l}^{(m)} + \beta_d^{(m)} \text{ for } 1 \leq d \leq p_m, \\ \delta_d^{(m)} &= \textstyle\sum_{l=1}^{\min(d,p_0)} \gamma_l \delta_{d-l}^{(m)} \text{ for } d > p_m. \end{aligned}\right\} \tag{3.3}$$

The coefficients $\delta_d^{(m)}$ represent what the effect on y_t is d periods after a non permanent change of one unit in $x_t^{(m)}$ occurred. Although $\delta^{(m)}(L)$ is generally of infinite order, this will not be the case when all roots of $\gamma(L)$ are roots of $\beta^{(m)}(L)$ too. For instance, if $p_0 = p_m$ and for $l \geq 1$ it happens to be the case that $\beta_l^{(m)} = \beta_0^{(m)} \gamma_l$, then $\delta^{(m)}(L) = \beta_0^{(m)}$. Hence, in this case all roots of the lag polynomials $\gamma(L)$ and $\beta^{(m)}(L)$ are similar and $\delta^{(m)}(L)$ is of order zero.

For general stable models (3.1) the immediate effect on $E(y_t)$ of a unit change in $\bar{x}^{(m)}$ is $\delta_0^{(m)} = \beta_0^{(m)}$. This is also called the **impact multiplier**. The effect after D periods of a permanent unit change in $\bar{x}^{(m)}$ is called the D^{th} **interim multiplier** and is given by

$$\delta_D^{*(m)} = \textstyle\sum_{d=0}^{D} \delta_d^{(m)} \text{ for } D = 0, 1, 2, ... \tag{3.4}$$

Note that $\delta_0^{*(m)} = \delta_0^{(m)}$ and $\delta_\infty^{*(m)} = TM_m$. The proportion of the total change completed after D periods is given by

$$\delta_D^{\dagger(m)} = \delta_D^{*(m)}/\delta_\infty^{*(m)}. \tag{3.5}$$

When all coefficients $\delta_d^{(m)}$ have the same sign, then the average or **mean lag** is defined as $\bar{D}^{(m)} = \sum_{d=0}^\infty d\delta_d^{(m)}/\delta_\infty^{*(m)}$. An alternative way to express this is as follows. Differentiating polynomial $\delta^{(m)}(L)$ with respect to L gives $\delta^{(m)\prime}(L) = \sum_{d=1}^\infty d\delta_d^{(m)}L^{d-1}$, hence $\bar{D}^{(m)} = \delta^{(m)\prime}(1)/\delta^{(m)}(1)$. Using $\delta^{(m)\prime}(L) = [\gamma(L)\beta^{(m)\prime}(L) - \gamma'(L)\beta^{(m)}(L)]/[\gamma(L)]^2$ we find

$$\bar{D}^{(m)} = \beta^{(m)\prime}(1)/\beta^{(m)}(1) - \gamma'(1)/\gamma(1). \tag{3.6}$$

The **median lag** $\ddot{D}^{(m)}$ is found by solving D from $\delta_D^{\dagger(m)} = 0.5$. It expresses how long it takes until 50% of the total multiplier has been realized. Note that when $\delta_d^{(m)}$ is oscillating there may be multiple solutions. All these characterizations of the dynamic effects of $x_t^{(m)}$ on y_t concern the hypothetical situation that all other exogenous regressors remain constant.

3.3. Particulars of low order ADL panel models

We return now to panel data models and consider first the very simple dynamic model given by

$$y_{i,t} = \gamma y_{i,t-1} + \beta^{(1)}x_{i,t}^{(1)} + \beta^{(2)}x_{i,t}^{(2)} + \eta_i + \varepsilon_{i,t}, \tag{3.7}$$

where $-1 < \gamma < 1$. This stable ADL(1,0,0) panel model is well known as a partial adjustment model. By successive substitution this can be rewritten as

$$\begin{aligned}y_{i,t} &= \beta^{(1)}x_{i,t}^{(1)} + \beta^{(2)}x_{i,t}^{(2)} + \gamma(\beta^{(1)}x_{i,t-1}^{(1)} + \beta^{(2)}x_{i,t-1}^{(2)} + \gamma y_{i,t-2} + \eta_i + \varepsilon_{i,t-1}) + \eta_i + \varepsilon_{i,t} \\ &= \beta^{(1)}x_{i,t}^{(1)} + \beta^{(2)}x_{i,t}^{(2)} + \gamma\beta^{(1)}x_{i,t-1}^{(1)} + \gamma\beta^{(2)}x_{i,t-1}^{(2)} + \gamma^2 y_{i,t-2} + (1+\gamma)\eta_i + \varepsilon_{i,t} + \gamma\varepsilon_{i,t-1} \\ &= \ldots.. \\ &= \beta^{(1)}\Sigma_{l=0}^\infty \gamma^l x_{i,t-l}^{(1)} + \beta^{(2)}\Sigma_{l=0}^\infty \gamma^l x_{i,t-l}^{(2)} + (1-\gamma)^{-1}\eta_i + \Sigma_{l=0}^\infty \gamma^l \varepsilon_{i,t-l}.\end{aligned}$$

The latter expression shows that the effects on $y_{i,t}$ of a permanent one unit change in the exogenous regressors $x_{i,t}^{(1)}$ or $x_{i,t}^{(2)}$ are in the long-run equal to $\beta^{(1)}/(1-\gamma)$ and $\beta^{(2)}/(1-\gamma)$, respectively. Moreover, the effects on $y_{i,t}$ of a non permanent one unit change in $x_{i,t}^{(1)}$ or $x_{i,t}^{(2)}$ (so, both $x_{i,t}^{(1)}$ and $x_{i,t}^{(2)}$ are assumed constant through time, but only at time period t_0 they increase by one unit) varies and are $\beta^{(1)}\gamma^l$ and $\beta^{(2)}\gamma^l$ respectively at time period t_0+l, for $l = 0, 1, 2, \ldots$. So, both long-run and short-run effects of $x_{i,t}^{(1)}$ on $y_{i,t}$, in relation to those of $x_{i,t}^{(2)}$, are similar, apart from a factor $\beta^{(1)}/\beta^{(2)}$. Hence, in some sense, the dynamic reaction patterns are parallel, since the effects decay exponentially at rate γ for both, although they start at the different levels $\beta^{(1)}$ and $\beta^{(2)}$ and next accumulate to the different magnitudes $\beta^{(1)}/(1-\gamma)$ and $\beta^{(2)}/(1-\gamma)$.

Now suppose the model contains the lags of $x_{i,t}^{(1)}$ and $x_{i,t}^{(2)}$ as well, hence we have

$$y_{i,t} = \gamma y_{i,t-1} + \beta_0^{(1)}x_{i,t}^{(1)} + \beta_1^{(1)}x_{i,t-1}^{(1)} + \beta_0^{(2)}x_{i,t}^{(2)} + \beta_1^{(2)}x_{i,t-1}^{(2)} + \eta_i + \varepsilon_{i,t}, \tag{3.8}$$

which is an ADL(1,1,1) panel model. Successive substitution yields now

$$\begin{aligned}
y_{i,t} &= \beta_0^{(1)}x_{i,t}^{(1)} + \beta_1^{(1)}x_{i,t-1}^{(1)} + \beta_0^{(2)}x_{i,t}^{(2)} + \beta_1^{(2)}x_{i,t-1}^{(2)} \\
&+ \gamma(\beta_0^{(1)}x_{i,t-1}^{(1)} + \beta_1^{(1)}x_{i,t-2}^{(1)} + \beta_0^{(2)}x_{i,t-1}^{(2)} + \beta_1^{(2)}x_{i,t-2}^{(2)} + \gamma y_{i,t-2} + \eta_i + \varepsilon_{i,t-1}) + \eta_i + \varepsilon_{i,t} \\
&= \beta_0^{(1)}x_{i,t}^{(1)} + (\beta_1^{(1)} + \gamma\beta_0^{(1)})x_{i,t-1}^{(1)} + \gamma\beta_1^{(1)}x_{i,t-2}^{(1)} + \beta_0^{(2)}x_{i,t}^{(2)} + (\beta_1^{(2)} + \gamma\beta_0^{(2)})x_{i,t-1}^{(2)} \\
&+ \gamma\beta_1^{(2)}x_{i,t-2}^{(2)} + \gamma^2 y_{i,t-2} + (1+\gamma)\eta_i + \varepsilon_{i,t} + \gamma\varepsilon_{i,t-1} \\
&= \beta_0^{(1)}x_{i,t}^{(1)} + (\beta_1^{(1)} + \gamma\beta_0^{(1)})x_{i,t-1}^{(1)} + \gamma\beta_1^{(1)}x_{i,t-2}^{(1)} + \beta_0^{(2)}x_{i,t}^{(2)} + (\beta_1^{(2)} + \gamma\beta_0^{(2)})x_{i,t-1}^{(2)} \\
&+ \gamma\beta_1^{(2)}x_{i,t-2}^{(2)} + \gamma^2(\beta_0^{(1)}x_{i,t-2}^{(1)} + \beta_1^{(1)}x_{i,t-3}^{(1)} + \beta_0^{(2)}x_{i,t-2}^{(2)} + \beta_1^{(2)}x_{i,t-3}^{(2)} + \gamma y_{i,t-3} + \eta_i + \varepsilon_{i,t-2}) \\
&+ (1+\gamma)\eta_i + \varepsilon_{i,t} + \gamma\varepsilon_{i,t-1} \\
&= \beta_0^{(1)}x_{i,t}^{(1)} + \Sigma_{l=1}^{\infty}\gamma^{l-1}(\gamma\beta_0^{(1)} + \beta_1^{(1)})x_{i,t-l}^{(1)} + \beta_0^{(2)}x_{i,t}^{(2)} + \Sigma_{l=1}^{\infty}\gamma^{l-1}(\gamma\beta_0^{(2)} + \beta_1^{(2)})x_{i,t-l}^{(2)} \\
&+ (1-\gamma)^{-1}\eta_i + \Sigma_{l=0}^{\infty}\gamma^l\varepsilon_{i,t-l}.
\end{aligned}$$

So, here the effects on $y_{i,t}$ of a permanent one unit change in $x_{i,t}^{(1)}$ or $x_{i,t}^{(2)}$ (the long-run multipliers) are found to be $\beta_0^{(1)} + (\gamma\beta_0^{(1)} + \beta_1^{(1)})(1-\gamma)^{-1} = (\beta_0^{(1)} + \beta_1^{(1)})/(1-\gamma)$ and $(\beta_0^{(2)} + \beta_1^{(2)})/(1-\gamma)$ respectively, which correslonds to (3.2). The impact multipliers are $\beta_0^{(1)}$ and $\beta_0^{(2)}$ respectively, and the impacts after one period are $\gamma\beta_0^{(1)} + \beta_1^{(1)}$ and $\gamma\beta_0^{(2)} + \beta_1^{(2)}$ respectively. These two are not necessarily parallel, because $\beta_0^{(1)}/\beta_0^{(2)} \neq (\gamma\beta_0^{(1)} + \beta_1^{(1)})/(\gamma\beta_0^{(2)} + \beta_1^{(2)})$ unless $\beta_0^{(1)}\beta_1^{(2)} = \beta_1^{(1)}\beta_0^{(2)}$. However, the effects of a non permanent one unit change are parallel from lag one onwards, because $\gamma^{l-1}(\gamma\beta_0^{(1)} + \beta_1^{(1)})/[\gamma^{l-1}(\gamma\beta_0^{(2)} + \beta_1^{(2)})] = (\gamma\beta_0^{(1)} + \beta_1^{(1)})/(\gamma\beta_0^{(2)} + \beta_1^{(2)})$. Hence, in this specification with two extra coefficients the ratio of the two impact multipliers $\beta_0^{(1)}/\beta_0^{(2)}$ is not necessarily equal to that of the long-run multipliers $(\beta_0^{(1)} + \beta_1^{(1)})/(\beta_0^{(2)} + \beta_1^{(2)})$, nor to the ratio of the impacts from lag one onwards. But, for both the lag patterns decay from lag one onwards at the same exponential rate γ.

There are a few particular cases of special interest. When $\beta_0^{(1)} = -\beta_1^{(1)}$ the long-run multiplier of $x^{(1)}$ with respect to y is zero, and hence $x^{(1)}$ has only temporary effects on y. When $\beta_1^{(1)} = -\gamma\beta_0^{(1)}$ then the impact multiplier of $x^{(1)}$ on y is equal to the total multiplier, namely $\beta_0^{(1)}$ and there are no delays in the effects of $x^{(1)}$ on y. When $\beta_1^{(2)} = -\gamma\beta_0^{(2)}$ too, model (3.8) specializes to

$$y_{i,t} - \gamma y_{i,t-1} = \beta_0^{(1)}(x_{i,t}^{(1)} - \gamma x_{i,t-1}^{(1)}) + \beta_0^{(2)}(x_{i,t}^{(2)} - \gamma x_{i,t-1}^{(2)}) + \eta_i + \varepsilon_{i,t}$$

or

$$y_{i,t} = \beta_0^{(1)}x_{i,t}^{(1)} + \beta_0^{(2)}x_{i,t}^{(2)} + (1-\gamma)^{-1}\eta_i + u_{i,t}, \qquad u_{i,t} = \gamma u_{i,t-1} + \varepsilon_{i,t}, \tag{3.9}$$

which is a panel ADL(0,0,0) model with AR(1) errors. This result indicates that an ADL model with AR(p) errors is just a special case of an ADL model with white-noise errors, where the orders of all lag polynomials have been increased by p, but where the lag coefficients obey particular nonlinear restrictions. This finding instigates to start our model specification search by including at least one lag of all regressors, because validity of internal instruments constructed from lagged not strictly exogenous regressors requires white-noise disturbances, and obtaining white noise disturbances is promoted by using sufficiently large orders of all lag polynomials.

When we add to the ADL(1,1,1) model $\beta_2^{(1)}x_{i,t-2}^{(1)}$ and $\beta_2^{(2)}x_{i,t-2}^{(2)}$ we obtain an ADL(1,2,2) specification. Then, apart from the ratio between the two long-run multipliers and that of the two impact multipliers, also the ratio of the effects after one period may have a unique value, but from lag two onwards there is decay at the same exponential rate determined by γ. Hence, basically this just leads to an extra subtlety in the short-run dynamics at lag 1. It should be obvious now what the effects would be of further increases of p_m, while keeping $p_0 = 1$.

However, if we rename $\gamma y_{i,t-1}$ in $\gamma_1 y_{i,t-1}$, and add $\gamma_2 y_{i,t-2}$ as well, whereas the values of γ_1 and γ_2 are such that the dynamic process is still stable, then in this ADL(2,2,2) model there is still a kind of parallel decay after one period, but not necessarily monotonically exponentially decreasing, but probably oscillating while gradually decreasing at the same time. Whereas on the other hand, when no lagged dependent regressors occur and $p_0 = 0$, then the dynamic patterns implied by changes in any of the $x^{(m)}$ are completely determined by the coefficients of the lag polynomials $\beta^{(m)}(L)$, which means that without further coefficient restrictions there is no parallel decay. For each $x^{(m)}$ the lag pattern is then completely determined by $\beta^{(m)}(L)$.

Presence of parallel decay for different regressors $x^{(m)}$ may at first sight establish a curious restriction which one might better avoid. However, allowing $p_0 > 0$, which induces occurrence of forms of parallel decay in the dynamic processes implied by ADL models, can be rationalized by the following behavior of economic agents. As an example we take the ADL(2,...,2) model

$$y_{i,t} = \gamma_1 y_{i,t-1} + \gamma_2 y_{i,t-2} + \Sigma_{m=1}^M(\beta_0^{(m)}x_{i,t}^{(m)} + \beta_1^{(m)}x_{i,t-1}^{(m)} + \beta_2^{(m)}x_{i,t-2}^{(m)}) + \eta_i + \varepsilon_{i,t},$$

which, when using $TM^{(m)} = (\beta_0^{(m)} + \beta_1^{(m)} + \beta_2^{(m)})/(1-\gamma_1-\gamma_2)$, can be rewritten as

$$\begin{aligned}\Delta y_{i,t} = {} & -\gamma_2\Delta y_{i,t-1} + \Sigma_{m=1}^M[\beta_0^{(m)}\Delta x_{i,t}^{(m)} - \beta_2^{(m)}\Delta x_{i,t-1}^{(m)}] \\ & + (\gamma_1+\gamma_2-1)[y_{i,t-1} - \Sigma_{m=1}^M TM^{(m)}x_{i,t-1}^{(m)} - \eta_i/(1-\gamma_1-\gamma_2)] + \varepsilon_{i,t}. \qquad (3.10)\end{aligned}$$

The latter form is called the error-correction or the equilibrium-correction form, see Hendry (1995). Neglecting for the moment the disturbances $\varepsilon_{i,t}$ and assuming $\Delta x_{i,t}^{(1)} = ... = \Delta x_{i,t}^{(M)} = 0$ over a long period, a stationary equilibrium where $y_{i,t} = \Sigma_{m=1}^M TM^{(m)}x_{i,t}^{(m)} + \eta_i/(1-\gamma_1-\gamma_2)$ will be attained eventually. Then the second factor in square brackets in (3.10) will be zero, like all other terms in that equation. If the factor $y_{i,t-1} - \Sigma_{m=1}^M TM^{(m)}x_{i,t-1}^{(m)} - \eta_i/(1-\gamma_1-\gamma_2)$, which expresses the long-run disequilibrium at time $t-1$, were non-zero, this will trigger a change in $y_{i,t}$ determined by the magnitudes of the disequilibrium and the disequilibrium correction factor $(\gamma_1+\gamma_2-1)$. In fact, also any actual current changes $\Delta x_{i,t}^{(m)}$ and passed changes $\Delta x_{i,t-1}^{(m)}$ and $\Delta y_{i,t-1}$, as well as chocks $\varepsilon_{i,t}$, all affect a change in $y_{i,t}$ according to (3.10). Shocks $\varepsilon_{i,t}$ and current and passed changes $\Delta x_{i,t}^{(m)}$ may in fact aggravate disequilibria. However, in the long-run these disequilibria will be overcome by the error-correction mechanism, because under stability $\gamma_1+\gamma_2-1 < 0$. So, if the disequilibrium is negative ($y_{i,t-1}$ too low) the third term of (3.10) is positive and will help to achieve that $\Delta y_{i,t}$ is positive, and the other way around for a positive disequilibrium. Hence, agents do react to short-run shocks, but aim to avoid a disequilibrium in the long-run. The latter brings about the parallel dynamics for different m after the more independent short-run reactions. In models where $p_0 = 0$ (no lagged-dependent variables as regressors) there is no overall reaction by the agents to disequilibria as such, but only to changes in individual $x_{i,t}^{(m)}$ variables. In case of partial adjustment, so when $p_0 = 1$ and $p_1 = ... = p_M = 0$, then

$$\Delta y_{i,t} = \Sigma_{m=1}^{M}\beta_0^{(m)}\Delta x_{i,t}^{(m)} + (\gamma_1 - 1)[y_{i,t-1} - \Sigma_{m=1}^{M}TM^{(m)}x_{i,t-1}^{(m)} - \eta_i/(1-\gamma_1)] + \varepsilon_{i,t}.$$

Here it are just the disequilibrium and the impact multipliers $\beta_0^{(m)}$ which determine the fully parallel reaction patterns, under the perhaps curious restriction that for each m the impact multiplier $\beta_0^{(m)}$ and the total multiplier $TM^{(m)}$ are such that the reaction $\gamma_1 - 1$ to a disequilibrium is equal to $-\beta_0^{(m)}/TM^{(m)}$, whereas when $p_m > 0$ the error-correction factor $(\gamma_1 - 1)$ and the impact and total multipliers are variation free. In such cases the first $p_m - 1$ intermediate impact multipliers may break away from the parallel decay patterns which will occur from lag p_m onwards in ADL models.

Obviously, by adding longer lags, more subtle lag patterns can be described by the extra parameters. However, all the above models imply lag patterns which are similar for all subjects i. They do not allow the total or intermediate multipliers to be different for different subjects. Such homogeneity may not always be realistic.

3.4. Interactions in dynamic panel models

A severe restriction on the standard dynamic linear panel data model considered so far is the homogeneity of all slope coefficients with respect to all individual subjects in the sample. Allowing for some heterogeneity could be realized by splitting the sample in a few subgroups, provided there is coefficient homogeneity within each subgroup. In the extreme this may multiply the number of parameters by the number of required subgroups. In case just some multipliers differ continuously or discretely with respect to particular characteristics of the subjects one should extend the model with interaction variables. It allows heterogeneity in the lag patterns and total multipliers for different subjects i for individual variables $x^{(m)}$ at the cost of remarkably few extra parameters, probably just one, whereas splitting the sample in G subgroups multiplies the number of coefficients by G.

Let us first consider the specification

$$y_{i,t} = \gamma y_{i,t-1} + \beta^{(1)}x_{i,t}^{(1)} + \beta^{(2)}x_{i,t}^{(2)} + \phi x_{i,t}^{(1)}x_{i,t}^{(2)} + \eta_i + \varepsilon_{i,t}, \qquad (3.11)$$

where $\phi x_{i,t}^{(1)}x_{i,t}^{(2)}$ establishes an interaction term. Assuming $x_{i,t}^{(2)} = \bar{x}_i^{(2)}$ is constant over time, this yields

$$\begin{aligned}
y_{i,t} &= (\beta^{(1)} + \phi\bar{x}_i^{(2)})x_{i,t}^{(1)} + \beta^{(2)}\bar{x}_i^{(2)} + \gamma y_{i,t-1} + \eta_i + \varepsilon_{i,t} \\
&= (\beta^{(1)} + \phi\bar{x}_i^{(2)})x_{i,t}^{(1)} + \beta^{(2)}\bar{x}_i^{(2)} \\
&+ \gamma[(\beta^{(1)} + \phi\bar{x}_i^{(2)})x_{i,t-1}^{(1)} + \beta^{(2)}\bar{x}_i^{(2)} + \gamma y_{i,t-2} + \eta_i + \varepsilon_{i,t-1}] + \eta_i + \varepsilon_{i,t} \\
&= (\beta^{(1)} + \phi\bar{x}_i^{(2)})\Sigma_{l=0}^{\infty}\gamma^l x_{i,t-l}^{(1)} + \beta^{(2)}(1-\gamma)^{-1}\bar{x}_i^{(2)} + (1-\gamma)^{-1}\eta_i + \Sigma_{l=0}^{\infty}\gamma^l\varepsilon_{i,t-l}.
\end{aligned}$$

Hence, the impact multiplier of $x_{i,t}^{(1)}$ with respect to $y_{i,t}$ is now $\beta^{(1)} + \phi\bar{x}_i^{(2)}$, and the total multiplier is $(\beta^{(1)} + \phi\bar{x}_i^{(2)})/(1-\gamma)$. These vary with the level of $\bar{x}_i^{(2)}$, whereas the lag pattern is still simply exponentially decaying. Similarly we find for the effect of $x_{i,t}^{(2)}$ that it has individual specific impact multiplier $\beta^{(2)}+\phi\bar{x}_i^{(1)}$, total multiplier $(\beta^{(2)}+\phi\bar{x}_i^{(1)})/(1-\gamma)$ and for each individual an exponential lag pattern with the same coefficient γ. Note that the lag patterns are again parallel over the whole range. If we had used $\phi x_{i,t-1}^{(1)}x_{i,t-1}^{(2)}$ as the only interaction term, then the immediate impacts would (again) be $\beta^{(1)}$ and $\beta^{(2)}$ (hence equal for all subjects i), but for time-constant $x_{i,t}^{(2)}=\bar{x}_i^{(2)}$ we obtain

$$\begin{aligned} y_{i,t} &= \beta^{(1)}x_{i,t}^{(1)}+\beta^{(2)}\bar{x}_i^{(2)}+\phi x_i^{(2)}x_{i,t-1}^{(1)} \\ &+\gamma(\beta^{(1)}x_{i,t-1}^{(1)}+\beta^{(2)}\bar{x}_i^{(2)}+\phi x_i^{(2)}x_{i,t-2}^{(1)}+\gamma y_{i,t-2}+\eta_i+\varepsilon_{i,t-1})+\eta_i+\varepsilon_{i,t} \\ &= \beta^{(1)}x_{i,t}^{(1)}+(\phi\bar{x}_i^{(2)}+\gamma\beta^{(1)})x_{i,t-1}^{(1)}+\gamma\phi\bar{x}_i^{(2)}x_{i,t-2}^{(1)}+(\beta^{(2)}+\gamma\beta^{(2)})\bar{x}_i^{(2)} \\ &+\gamma^2 y_{i,t-2}+(1+\gamma)\eta_i+\varepsilon_{i,t}+\gamma\varepsilon_{i,t-1} \\ &= \beta^{(1)}x_{i,t}^{(1)}+(\beta^{(1)}\gamma+\phi\bar{x}_i^{(2)})\Sigma_{l=1}^{\infty}\gamma^{l-1}x_{i,t-l}^{(1)}+\beta^{(2)}(1-\gamma)^{-1}\bar{x}_i^{(2)} \\ &+(1-\gamma)^{-1}\eta_i+\Sigma_{l=0}^{\infty}\gamma^l\varepsilon_{i,t-l}. \end{aligned} \tag{3.12}$$

So the total multiplier is again $\beta^{(1)}+(\beta^{(1)}\gamma+\phi\bar{x}_i^{(2)})/(1-\gamma)=(\beta^{(1)}+\phi\bar{x}_i^{(2)})/(1-\gamma)$, but now the exponential decay only starts from lag one onwards. Of course one could also include both interactions $\phi_0 x_{i,t}^{(1)}x_{i,t}^{(2)}$ and $\phi_1 x_{i,t-1}^{(1)}x_{i,t-1}^{(2)}$. Then also the immediate impact varies per subject, whereas the effect of $\bar{x}_i^{(2)}$ on the impacts from lag one onwards may be different.

Next consider the same options for adding interactions to the model with general first-order dynamics instead of the just considered specific partial-adjustment case. First we examine

$$y_{i,t}=\gamma y_{i,t-1}+\beta_0^{(1)}x_{i,t}^{(1)}+\beta_1^{(1)}x_{i,t-1}^{(1)}+\beta_0^{(2)}x_{i,t}^{(2)}+\beta_1^{(2)}x_{i,t-1}^{(2)}+\phi x_{i,t}^{(1)}x_{i,t}^{(2)}+\eta_i+\varepsilon_{i,t}. \tag{3.13}$$

This yields for constant $x_{i,t}^{(2)}$

$$y_{i,t}=\gamma y_{i,t-1}+(\beta_0^{(1)}+\phi\bar{x}_i^{(2)})x_{i,t}^{(1)}+\beta_1^{(1)}x_{i,t-1}^{(1)}+(\beta_0^{(2)}+\beta_1^{(2)})\bar{x}_i^{(2)}+\eta_i+\varepsilon_{i,t},$$

which has similarities with (3.8) so that we can directly indicate that the long-run multiplier is $(\beta_0^{(1)}+\beta_1^{(1)}+\phi\bar{x}_i^{(2)})/(1-\gamma)$, the impact multiplier is $\beta_0^{(1)}+\phi\bar{x}_i^{(2)}$ and from lag one onwards exponential decay sets in with the impacts equal to $\gamma^l(\beta_0^{(1)}+\phi\bar{x}_i^{(2)})+\gamma^{l-1}\beta_1^{(1)}$. The other option, where we include $\phi x_{i,t-1}^{(1)}x_{i,t-1}^{(2)}$, gives

$$y_{i,t}=\gamma y_{i,t-1}+\beta_0^{(1)}x_{i,t}^{(1)}+(\beta_1^{(1)}+\phi\bar{x}_i^{(2)})x_{i,t-1}^{(1)}+(\beta_0^{(2)}+\beta_1^{(2)})\bar{x}_i^{(2)}+\eta_i+\varepsilon_{i,t},$$

which yields the same total multiplier, a constant impact $\beta_0^{(1)}$ and next interim impacts equal to $\gamma^l\beta_0^{(1)}+\gamma^{l-1}(\beta_1^{(1)}+\phi\bar{x}_i^{(2)})$. Including both the unlagged and lagged interactions yields

$$y_{i,t}=\gamma y_{i,t-1}+(\beta_0^{(1)}+\phi_0\bar{x}_i^{(2)})x_{i,t}^{(1)}+(\beta_1^{(1)}+\phi_1\bar{x}_i^{(2)})x_{i,t-1}^{(1)}+(\beta_0^{(2)}+\beta_1^{(2)})\bar{x}_i^{(2)}+\eta_i+\varepsilon_{i,t},$$

giving total multiplier $[\beta_0^{(1)}+\beta_1^{(1)}+\bar{x}_i^{(2)}(\phi_0+\phi_1)]/(1-\gamma)$, impact $\beta_0^{(1)}+\phi_0\bar{x}_i^{(2)}$, followed by interim impacts equal to $\gamma^l(\beta_0^{(1)}+\phi_0\bar{x}_i^{(2)})+\gamma^{l-1}(\beta_1^{(1)}+\phi_1\bar{x}_i^{(2)})=\gamma^{l-1}[\gamma\beta_0^{(1)}+\beta_1^{(1)}+(\gamma\phi_0+\phi_1)\bar{x}_i^{(2)}]$. The latter has, of course, the earlier two as special cases.

Possibly in many actual cases specification (3.13) already provides sufficient flexibility to model variability per subject of the total multipliers and lag patterns in both variables $x_{i,t}^{(1)}$ and $x_{i,t}^{(2)}$. Of course, when models contain more explanatories than just $x_{i,t}^{(1)}$ and $x_{i,t}^{(2)}$ further interactions can be introduced leading to variability of impacts with respect to more than just one other regressor.

Yet another option is to involve the lagged dependent variable itself in the interactions. We examine the case

$$y_{i,t} = \gamma y_{i,t-1} + \phi x_{i,t-1}^{(2)} y_{i,t-1} + \beta_0^{(1)} x_{i,t}^{(1)} + \beta_1^{(1)} x_{i,t-1}^{(1)} + \beta_0^{(2)} x_{i,t}^{(2)} + \beta_1^{(2)} x_{i,t-1}^{(2)} + \eta_i + \varepsilon_{i,t}.$$

For constant $x_{i,t}^{(2)}$ this implies

$$y_{i,t} = (\gamma + \phi \bar{x}_i^{(2)}) y_{i,t-1} + \beta_0^{(1)} x_{i,t}^{(1)} + \beta_1^{(1)} x_{i,t-1}^{(1)} + (\beta_0^{(2)} + \beta_1^{(2)}) \bar{x}_i^{(2)} + \eta_i + \varepsilon_{i,t}.$$

Now the requirement that the lagged dependent variable coefficient is smaller than 1 in absolute value in order to have sensible behavioral stable dynamic processes is at challenge. Therefore, only for particular $x_{i,t}^{(2)}$ variables such interactions seem manageable. In particular when $x_{i,t}^{(2)}$ is a dummy variable it enables to distinguish different coefficient values of the lagged dependent variable between two subgroups.

From the above we conclude regarding relaxing the homogeneity of all subjects with respect to the total impact of the separate determining factors and their dynamic patterns that it seems promising in many cases to just augment the standard dynamic panel data specification by the products of couples of current regressors. Note that when $M = 10$ this already implies the addition of possibly 55 interaction terms, which then also includes the squares of the regressors $x_{i,t}^{(m)}$ to allow for a parabolic relationship. Even more subtle heterogeneity can be parameterized by including the products of triples of regressors.

3.5. Dynamic multipliers in panel models with feedbacks

Let us now consider the situation where in model (3.1) the variables $x_{i,t}^{(m)}$ are exogenous for $m > 1$, whereas $x_{i,t}^{(1)}$ is predetermined, and $E(x_{i,t-l}^{(1)} \varepsilon_{i,t}) = 0$ and $E(x_{i,t+1+l}^{(1)} \varepsilon_{i,t}) \neq 0$ for $l \geq 0$. We will first examine a relatively simple case, namely

$$x_{i,t}^{(1)} = \phi \varepsilon_{i,t-1} + \psi y_{i,t-1} + \ddot{x}_{i,t}^{(1)}, \tag{3.14}$$

where $\ddot{x}_{i,t}^{(1)}$ is the joint contribution of all other determinants of $x_{i,t}^{(1)}$, which are all exogenous with respect to $\varepsilon_{i,t}$, hence $E(\varepsilon_{i,t} \ddot{x}_{i,s}^{(1)}) = 0 \; \forall t, s$. Note that $x_{i,t}^{(1)}$ is affected by lagged feedback from $\varepsilon_{i,t}$, directly if $\phi \neq 0$ and also indirectly if $\psi \neq 0$.

Now consider the hypothetical situation in which $\ddot{x}_{i,t}^{(1)}$ and all exogenous regressors of $y_{i,t}$ are running at constant values. Then $E(x_{i,t}^{(1)}) = \bar{x}_i^{(1)}$ and $E(y_{i,t}) = \bar{y}_i$ will remain constant too. A permanent increase of $\bar{x}_i^{(1)}$ at time period t_0 by one unit, due to an increase by one unit of $\ddot{x}_{i,t}^{(1)}$ at t_0, whereas all other exogenous variables remain constant, will have the following effects. Although $\Delta \bar{x}_{i,t}^{(1)} = \bar{x}_{i,t}^{(1)} - \bar{x}_{i,t-1}^{(1)} = 0$ for $t < t_0$, we have $\Delta \bar{x}_{i,t_0}^{(1)} = 1$, whereas $\Delta \bar{x}_{i,t}^{(1)} = \psi \Delta \bar{y}_{i,t-1}$ for $t > t_0$, while $\Delta x_{i,t}^{(m)} = 0$ for $m = 2, ..., M$ and $\forall t$. Since $\gamma(L) \bar{y}_{i,t} = \sum_{m=1}^{M} \beta^{(m)}(L) \bar{x}_{i,t}^{(m)}$, it follows that $\gamma(L) \Delta \bar{y}_{i,t} = \beta^{(1)}(L) \Delta \bar{x}_{i,t}^{(1)}$. This yields $\Delta\bar{y}_{i,t_0} = \beta_0^{(1)}\Delta x_{i,t_0}^{(1)} = \beta_0^{(1)}$, because $\Delta\bar{y}_{i,t} = 0$, $\Delta\bar{x}_{i,t}^{(1)} = 0$ for $t < t_0$ and $\Delta\bar{x}_{i,t_0}^{(1)} = 1$. So, the immediate effect on $y_{i,t}$ is $\beta_0^{(1)}$. It also yields

$$\begin{aligned}\Delta\bar{y}_{i,t_0+1} &= \gamma_1\Delta\bar{y}_{i,t_0} + \beta_0^{(1)}\Delta\bar{x}_{i,t_0+1}^{(1)} + \beta_1^{(1)}\Delta\bar{x}_{i,t_0}^{(1)} \\ &= (\gamma_1 + \beta_0^{(1)}\psi)\Delta\bar{y}_{i,t_0} + \beta_1^{(1)} = (\gamma_1 + \beta_0^{(1)}\psi)\beta_0^{(1)} + \beta_1^{(1)}. \end{aligned} \tag{3.15}$$

This illustrates that all the interim multipliers (apart from the impact multiplier) differ from the formulas derived earlier in case $\psi \neq 0$. For $\psi = 0$ we have $\Delta\bar{x}_{i,t}^{(1)} = 0$ for $t > t_0$, so $\Delta\bar{y}_{i,t_0+1} = \gamma_1\beta_0^{(1)} + \beta_1^{(1)}$ as in the case of exogenous $x_{i,t}^{(1)}$. Apparently the value of ϕ is irrelevant.

More generally, we may face the situation where

$$\pi(L)x_{i,t}^{(1)} = \phi(L)\varepsilon_{i,t} + \psi(L)y_{i,t} + \ddot{x}_{i,t}^{(1)}, \tag{3.16}$$

with $\pi(L) = 1 - \pi_1 L - ... - \pi_{p_\pi}L^{p_\pi}$, $\phi(L) = \phi_0 + \phi_1 L + ... + \phi_{p_\phi}L^{p_\phi}$ and $\psi(L) = \psi_0 + \psi_1 L + ... + \psi_{p_\psi}L^{p_\psi}$, with the integer orders p_π, p_ϕ and p_ψ all nonnegative. Assuming again that $\ddot{x}_{i,t}^{(1)}$ and $x_{i,t}^{(m)}$ for $m > 1$ are all exogenous, $x_{i,t}^{(1)}$ is an endogenous regressor in our model of primary interest if either $\phi_0 \neq 0$ or $\psi_0 \neq 0$ or both, whereas regressor $x_{i,t}^{(1)}$ is exogenous if $\phi(L) = 0$ and $\psi(L) = 0$ and it is predetermined in all other cases. Taking into account that the single relationship under study may contain several endogenous and predetermined regressors, which probably are determinants of $x_{i,t}^{(1)}$ too, it is obvious that to be able to characterize the dynamic effects of a change in a predetermined or endogenous regressor on dependent variable $y_{i,t}$ it appears required to specify and analyze a whole system of structural equations, instead of just one single equation. Then by substitution the so-called final form equation for $y_{i,t}$ can be obtained, in which $y_{i,t}$ is expressed in just its lags, (lags of) all the exogenous variables of the system and all disturbances of the system. In most practical situations finding the proper specification of the final form equation seems extremely difficult, if not impossible. Therefore, a more realistic and pragmatic approach may be, to analyze dynamic multipliers as indicated in the preceding subsections, thus neglecting any possible indirect feedbacks via endogenous an predetermined regressors $x_{i,t}^{(m)}$, though explicitly mentioning that the obtained results concern the primary dynamic effects, upon neglecting any secondary effects. Of course, such calculations are of modest usefulness, not in the least because supplementing them with a meaningful assessment of their accuracy by confidence intervals is generally beyond reach.

3.6. Primary interim multipliers in the presence of interactions

Let the finally accepted estimated dynamic model include lags of squared regressors and of interactions between couples of the M explanatories $x_{i,t}^{(m)}$ and be given by

$$\begin{aligned}\Delta y_{i,t} = {}& \textstyle\sum_{l=1}^{p}\hat{\gamma}_l\Delta y_{i,t-l} + \sum_{m=1}^{M}\sum_{l=0}^{p}\hat{\beta}_l^{(m)}\Delta x_{i,t-l}^{(m)} + \sum_{m=1}^{M}\sum_{j=m}^{M}\sum_{l=0}^{p}\hat{\beta}_l^{(m,j)}\Delta(x_{i,t-l}^{(m)}x_{i,t-l}^{(j)}) \\ & + \textstyle\sum_{s=2}^{T}\hat{\tau}_s\Delta d_{i,t}^{(s)} + \Delta\hat{\varepsilon}_{i,t}. \end{aligned} \tag{3.17}$$

Here p is the maximum lag length that occurs. For the sake of simplicity we will assume $p = 2$ below. Due to the accepted imposed restrictions many of the coefficients $\hat{\gamma}_l$,

$\hat{\beta}_l^{(m)}$ and $\hat{\beta}_l^{(m,j)}$ may actually be zero, or obey other types of linear restrictions. We will derive now an algorithm for calculating the primary interim multipliers (neglecting other feedbacks than the direct feedbacks via $\hat{\gamma}_l$) for some hypothetical individual, say Q, of y_{Q,T^*+h} with respect to a small hypothetical one-off change at time period T^* in one of the M explanatory variables, say $x_{Q,T^*}^{(F)}$, where $1 \leq F \leq M$. These multipliers will be denoted as $IM_{Q,h}^{(F)}$, for $h = 0, ..., H$, where H is the time horizon. Hence, $IM_{Q,0}^{(F)}$ is the immediate multiplier and $IM_{Q,\infty}^{(F)}$ the total multiplier with respect to regressor F, provided dynamic process (3.17) is stable. Whereas in absence of interactions such multipliers, as presented in section 3.3, are invariant regarding the particular individual F and time period T^* considered, in presence of interactions the multipliers are individual specific and they depend on the actual state of the dynamic process for F over the period $T^* - p$ until T^*.

To find an expression for $IM_{Q,h}^{(F)}$ in terms of the relevant characteristics of individual Q and the coefficient estimates $\hat{\gamma}_l$, $\hat{\beta}_l^{(m)}$ and $\hat{\beta}_l^{(m,j)}$, we will first define a baseline projection for the dependent variable, indicated by $y_{Q,T^*+h}^{(0)}$, and next M level shift projections, indicated by $y_{Q,T^*+h}^{(F)}$, for $F = 1, ..., M$, for which the baseline value $x_{Q,T^*}^{(F)}$ is replaced by $(1+q)x_{Q,T^*}^{(F)}$, with q some small real value. From these projections the M series of interim multipliers can be obtained. They are given by

$$IM_{Q,h}^{(F)} = (y_{Q,T^*+h}^{(F)} - y_{Q,T^*+h}^{(0)})/qx_{Q,T^*}^{(F)}, \; h = 0, ..., H. \tag{3.18}$$

For q not larger than, say, 0.1 in absolute value, we expect very modest dependence of $IM_{Q,h}^{(F)}$ on q.

For the baseline projection we assume that $\Delta x_{Q,T^*+h}^{(m)} = \Delta \hat{\varepsilon}_{Q,T^*+h} = 0$ for $h > 0$ and all m. This implies

$$\left.\begin{aligned}
\Delta y_{Q,T^*}^{(0)} &= \Delta y_{Q,T^*}, \\
\Delta y_{Q,T^*+1}^{(0)} &= \hat{\gamma}_1 \Delta y_{Q,T^*} + \hat{\gamma}_2 \Delta y_{Q,T^*-1} + \textstyle\sum_{m=1}^{M} \sum_{l=0}^{1} \hat{\beta}_{l+1}^{(m)} \Delta x_{i,T^*-l}^{(m)} \\
&\qquad + \textstyle\sum_{m=1}^{M} \sum_{j=m}^{M} \sum_{l=0}^{1} \hat{\beta}_{l+1}^{(m,j)} \Delta (x_{i,T^*-l}^{(m)} x_{i,T^*-l}^{(j)}), \\
\Delta y_{Q,T^*+2}^{(0)} &= \hat{\gamma}_1 \Delta y_{Q,T^*+1}^{(0)} + \hat{\gamma}_2 \Delta y_{Q,T^*} + \textstyle\sum_{m=1}^{M} \hat{\beta}_{2}^{(m)} \Delta x_{i,T^*}^{(m)} \\
&\qquad + \textstyle\sum_{m=1}^{M} \sum_{j=m}^{M} \hat{\beta}_{2}^{(m,j)} \Delta (x_{i,T^*}^{(m)} x_{i,T^*}^{(j)}), \\
\Delta y_{Q,T^*+h}^{(0)} &= \hat{\gamma}_1 \Delta y_{Q,T^*+h-1}^{(0)} + \hat{\gamma}_2 \Delta y_{Q,T^*+h-2}^{(0)}, \; h = 3, ..., H.
\end{aligned}\right\} \tag{3.19}$$

For the level shift projection with respect to the F^{th} explanatory variable there is just one different underlying assumption, namely that $x_{Q,T^*}^{(F)}$ is multiplied by $1+q$. So, $\Delta x_{Q,T^*-l}^{(m)}$ and $\Delta(x_{Q,T^*-l}^{(m)} x_{Q,T^*-l}^{(j)})$, for both m and j different from F, do not change, nor does $\Delta x_{Q,T^*-l}^{(F)}$ for $l > 0$, but $\Delta x_{Q,T^*}^{(F)}$ should be replaced by $\Delta x_{Q,T^*}^{(F)} + qx_{Q,T^*}^{(F)}$, and for $m \neq F$ expression $\Delta(x_{Q,T^*}^{(F)} x_{Q,T^*}^{(m)}) = x_{Q,T^*}^{(F)} x_{Q,T^*}^{(m)} - x_{Q,T^*-1}^{(F)} x_{Q,T^*-1}^{(m)}$ should be replaced by $\Delta(x_{Q,T^*}^{(F)} x_{Q,T^*}^{(m)}) + qx_{Q,T^*}^{(F)} x_{Q,T^*}^{(m)}$, whereas $\Delta(x_{Q,T^*}^{(F)} x_{Q,T^*}^{(F)}) = x_{Q,T^*}^{(F)} x_{Q,T^*}^{(F)} - x_{Q,T^*-1}^{(F)} x_{Q,T^*-1}^{(F)}$ should be replaced by $\Delta(x_{Q,T^*}^{(F)} x_{Q,T^*}^{(F)}) + (2q + q^2) x_{Q,T^*}^{(F)} x_{Q,T^*}^{(F)}$. Hence, defining $\Delta x_{Q,T^*-l}^{*(m)}$ and $\Delta(x_{Q,T^*-l}^{*(m)} x_{Q,T^*-l}^{*(j)})$ such that they are equivalent to $\Delta x_{Q,T^*-l}^{(m)}$ and $\Delta(x_{Q,T^*-l}^{(m)} x_{Q,T^*-l}^{(j)})$ respectively, except for the replacements just mentioned, then the level shift projection can be expressed as

$$\left.\begin{array}{l}
\Delta y^{(F)}_{Q,T^*} = \Delta y_{Q,T^*} + \hat{\beta}^{(F)}_0 q x^{(F)}_{i,T^*} \\
\qquad\qquad +\sum_{m=1}^{M} \hat{\beta}^{(m,F)}_0 [\Delta(x^{*(F)}_{Q,T^*} x^{*(m)}_{Q,T^*}) - \Delta(x^{(F)}_{Q,T^*} x^{(m)}_{Q,T^*})], \\
\Delta y^{(F)}_{Q,T^*+1} = \hat{\gamma}_1 \Delta y^{(F)}_{Q,T^*} + \hat{\gamma}_2 \Delta y_{Q,T^*-1} + \sum_{m=1}^{M} \sum_{l=0}^{1} \hat{\beta}^{(m)}_{l+1} \Delta x^{*(m)}_{Q,T^*-l} \\
\qquad\qquad +\sum_{m=1}^{M} \sum_{j=m}^{M} \sum_{l=0}^{1} \hat{\beta}^{(m,j)}_{l+1} \Delta(x^{*(m)}_{Q,T^*-l} x^{*(j)}_{Q,T^*-l}), \\
\Delta y^{(F)}_{Q,T^*+2} = \hat{\gamma}_1 \Delta y^{(F)}_{Q,T^*+1} + \hat{\gamma}_2 \Delta y^{(F)}_{Q,T^*} + \sum_{m=1}^{M} \hat{\beta}^{(m)}_2 \Delta x^{*(m)}_{i,T^*} \\
\qquad\qquad +\sum_{m=1}^{M} \sum_{j=m}^{M} \hat{\beta}^{(m,j)}_2 \Delta(x^{*(m)}_{Q,T^*} x^{*(j)}_{Q,T^*}) \\
\Delta y^{(F)}_{Q,T^*+h} = \hat{\gamma}_1 \Delta y^{(F)}_{Q,T^*+h-1} + \hat{\gamma}_2 \Delta y^{(F)}_{Q,T^*+h-2},\ h = 3, ..., H.
\end{array}\right\} \tag{3.20}$$

From the above we find that the effect of the level-shift on the dependent variable after $h = 0, ..., H$ periods is given by

$$\begin{aligned}
y^{(F)}_{Q,T^*+h} - y^{(0)}_{Q,T^*+h} &= \Sigma_{l=0}^{h} \Delta y^{(F)}_{Q,T^*+l} + y_{Q,T^*-1} - \Sigma_{l=0}^{h} \Delta y^{(0)}_{Q,T^*+l} - y_{Q,T^*-1} \\
&= \Sigma_{l=0}^{h} (\Delta y^{(F)}_{Q,T^*+l} - \Delta y^{(0)}_{Q,T^*+l}).
\end{aligned}$$

Substitution of (3.19) and (3.20) in this expression enables to evaluate all the M series of interim multipliers (3.18).

Different series of multipliers will be obtained for different values of $x^{(m)}_{Q,T^*}, \Delta x^{(m)}_{Q,T^*}$ and $\Delta x^{(m)}_{Q,T^*-1}$ for $m = 1, ..., M$; it can easily be shown that the multipliers are in fact invariant with respect to $\Delta y_{Q.T^*}$ and $\Delta y_{Q.T^*-1}$. A straight-forward choice would be $x^{(m)}_{Q,T^*-l} = N^{-1} \Sigma_{i=1}^{N} x^{(m)}_{i,T-l}$ for $l = 0, 1, 2$. But, instead of focussing on the average individual, multipliers can also be calculated for more atypical individuals Q, perhaps by adding/subtracting a multiple of the value of the sample standard deviations of $x^{(m)}_{i,T-l}$ over the N subjects. If the model does not include interactions nor squared variables, then the interactions calculated according (3.18) will be equal to those of section 3.3 when one assumes $\Delta x^{(m)}_{Q,T^*} = \Delta x^{(m)}_{Q,T^*-1} = 0$.

4. Further practical considerations and a tentative strategy

Let us now try to integrate all issues discussed above into a practical strategy and set of instructions that can be implemented in Stata software via the package Xtabond2. In a first subsection we will address some of the peculiarities of this software, also in relation to a few general observations on what is presently known about the qualities of the various techniques in finite samples, for which we refer to the second half of Kiviet et al. (2017). Next we will develop a 10-stage practical modelling strategy for microeconometric single relationships.

4.1. Some unpleasant peculiarities

All established theoretical qualities of the econometric tools to be used are asymptotic in nature, namely for samples containing an infinite number of subjects (individuals, firms) over a short time span (N extremely large, T finite). In empirical samples, where N is finite too, depending on peculiarities still hardly understood, the actual qualities of the tools may differ, sometimes in a counterintuitive way, from their asymptotic qualities. Examples of such counterintuitive phenomena are: (i) 2-step GMM estimation, which employs 1-step residuals to take possible heteroskedasticity into account, although in theory more efficient than robust[9] 1-step GMM estimation, may actually be less accurate (produce estimates with larger mean squared errors), probably due to vulnerability regarding the estimated weighting matrix; (ii) discarding the use of some valid (though relatively weak) instruments, either by collapsing or curtailing the instrument matrix, although harmful for the efficiency of the limiting distribution of consistent GMM estimators, may actually reduce estimator bias in finite samples and thus yield improved accuracy.

As far as I know, for no serious finite sample problem in the analysis of dynamic panel data models a universally effective correction to overcome it has been developed yet. To a limited degree there is one prominent exception to this. That is the Windmeijer (2005) correction to the usually overoptimistic (negatively biased) standard errors produced by standard 2-step GMM. One should always employ this correction[10], though not overestimate its effectiveness, because often the true standard deviations are still substantially larger than the corrected ones. Moreover, the 2-step GMM coefficient estimates are often seriously biased too. Hence, just correcting the bias of the variance estimate in Wald-type coefficient tests and not the bias of the coefficients themselves may still lead to poor (or even worse) test performance.

With respect to instrument validity tests the following should be mentioned. All Hansen tests produced by Xtabond2 are obtained as a quadratic form in the product of the transposed instruments and the 2-step residuals, whereas the estimated covariance matrix of this vector has been obtained by employing 1-step instead of 2-step residuals. Therefore we will address it as $J^{(2,1)}$ like in Kiviet et al. (2017, p.18), where it has been compared with the variant suggested in Arellano and Bond (1991), which uses 2-step residuals for the covariance as well.

As suggested by Hayashi (2000, p.222), sub-optimal weighting is being used in Xtabond2 for incremental Hansen tests. This avoids negative outcomes of the test statistic. These may occur when one simply takes the difference of two standard Hansen tests. However, negative test outcomes as an approximation to an asymptotically non-negative statistic do not seem a major problem, as they simply suggest an extremely high p-value. Using sub-optimal weights, though, may sacrifice test power and possibly disturb size control. By calculating in an incremental Hansen test the overall Hansen test based on the restrained set of instruments while weighting is based on using the unrestrained set its distribution is shifted to the right. Apart from preventing negative outcomes, this may also affect the distribution in the right-hand tail. This could offset possible underrejection by a standard difference Hansen test[11], but could also lead to more serious overrejection. This is one of the many finite sample issues that has not

[9] The robustification here is with respect to estimating the variance matrix of the 1-step GMM coefficient estimates such that it is consistent under unknown forms of possible heteroskedasticity of the disturbances (either over individuals or over time or both). Robustness of variance estimates with respect to unknown forms of serial correlation is unachievable when some regressors are not exogenous, because serial correlation would render instruments invalid and GMM coefficient estimates inconsistent.

[10] In Xtabond 2 this correction is wrongly addressed as "robustifying" 2-step estimation. The correction produces a better approximation to the true standard errors than the standard asymptotic one, by employing higher-order asymptotic methods. Both the uncorrected and the corrected version are robust with respect to unknown forms of heteroskedasticity.

[11] This can easily be obtained from Xtabond2 by subtracting manually two optimally weighted standard Hansen tests.

been thoroughly examined yet.

In summary, for producing inference on dynamic panel data models it is in general not clear yet which of the many alternative implementation options one should prefer in practice for test procedures and for estimators of coefficients and of their variance. The particular options provided by Xtabond2 are not always the only ones possible and probably not always the ones one should prefer.[12]

4.2. A sequential strategy for implementing GMM for micro panels

As a consequence of all the above mentioned obstacles, finding an initial candidate MSM will be far from easy, if not sheer impossible. This MSM should ideally be such that the true underlying data generating process of the structural relationship can ultimately be represented and accurately estimated after imposing the right coefficient restrictions and finding and exploiting any additional valid and effective moment conditions. Before we can get to this deductive specialization phase of imposing extra coefficient restrictions and exploiting extra moment conditions, we first need to design an inductive discovery phase to search for and lastly find an acceptable MSM. The latter phase is mainly conducted on the basis of the interpretation of p-values of a well-designed series of misspecification tests. These should reveal impermissible patterns of serial correlation in the disturbances, any omitted relevant regressors and invalidity of subsets of the exploited instruments. Next, after an acceptable MSM has been discovered, in the deductive phase the p-values of series of mostly very similar test procedures are being used. But now these tests are designed to reveal any redundant regressors or additional valid instruments to enable to achieve higher levels of efficiency.

Against the background provided above, and given the currently available still limited understanding of the performance in finite samples of the various statistical tools, the methodology that we want to recommend here embarks on a strategy consisting of the following ten sequential stages. All mentioned thresholds for p-values should be taken with substantial pinches of salt and be adapted to the actual situation regarding size of the available panel data set and any trustworthy expert knowledge regarding the relationship under study. Examples of the precise formulation of the required Xtabond2 code can be found in Appendix B. The 10 stages are:

1. Start with a specification of the model and instrument set which seems a reasonable compromise between generality in order to aim for consistency and specificity in order to acquire some precision as well. So, avoid to impose patently false restrictions on the model by omitting possibly relevant direct determinants of the dependent variable. Nor exploit any instruments that could well be invalid. Hence, include sufficient lags of all variables $x_{i,t}^{(m)}$ that seem relevant and some lags of $y_{i,t}$ too, not only to allow for a sufficiently rich pattern of the dynamic adjustments, but also to avoid serial correlation of the disturbances.[13] Treat all contemporaneous variables as endogenous with respect to $\varepsilon_{i,t}$, except when their exogeneity is

[12] In addition, in the current version the figure used for the number of regressors does not correct for any regressors omitted by the program (to avoid extreme multicollinearity, which will occur when too many time dummies are being included). As a consequence, the number of degrees of freedom of Sargan and Hansen tests may be deflated, and so will then be the associated asymptotic p-values.

[13] In many applied studies it seems to be believed wrongly that by adopting a partial adjustment framework, thus just including $y_{i,t-1}$ as the one and only lagged regressor, by its single extra coefficient,

beyond doubt. Though, at the same time, avoid to include regressors that most probably will actually be redundant or may jeopardize identification of the structural equation of interest for variable $y_{i,t}$, because their effect is not direct but just indirect via other possibly endogenous explanatories. Right from the beginning exploit a fair number of instruments which seem reasonably strong and should be valid if the chosen dynamic specification is general enough to avoid serial correlation. Because we will estimate the model after taking first-differences in order to remove time-constant unobserved individual effects, it is useless to include in the model any variables which are systematically time-constant for each unit over the whole time period covered by the sample, such as the intercept, etnicity, country of birth of individuals, or as may occur for gender, domicile, highest level of education, etc. Effects of such variables are absorbed by the individual effects. Include time dummies and use them as instruments too. To avoid the dummy-trap in (2.1) it should contain next to the individual effects just $T-1$ time-dummies, for instance for $t = 2, ..., T$. Their inclusion implies that explanatories which are a function of these should not be included too. This regards a linear (or quadratic, hence any systematic) trend, or the age of unit i (if the data have been collected on the same date each year) or inflation as experienced by all units, etc. Make sure that any scaling or other transformations used for the variables $y_{i,t}$ and $x_{i,t}^{(m)}$ for $m = 1, ..., M$ do make sense, so make a deliberate choice between including variables in real or in nominal terms, taking them as fractions or not, including them directly as observed or taking their logarithm, etc. The crucial issue is here: what transformation of the variables realizes that the coefficient of each regressor in the estimated model will really be constant in the sample and over the whole population. Hence, at stake is the constancy of the partial derivatives, or perhaps of the elasticities (which would require taking logs) or yet another characteristic, of the explanatory variables with respect to the dependent variable as observed in the sample. If these partial derivatives should vary with one or more other variables, this may require to include interaction terms.

2. Estimate this initial candidate MSM, taken in first-difference form, both by 1-step Arellano-Bond GMM, requesting heteroskedasticity robust standard errors, and by 2-step Arellano-Bond GMM with Windmeijer correction of its standard errors. For both 1-step and 2-step use the same set of instruments and formulate these such that separate incremental Hansen tests will be calculated for all the instruments obtained from each separate variable $x_{i,t}^{(m)}$ and from $y_{i,t}$. If the size of the sample permits, employ all the available internal instruments. Hence, use to construct the instruments for all endogenous (with respect to $\varepsilon_{i,t}$) regressors $x_{i,t}^{(m)}$, as well as for $y_{i,t}$, the option gmm(2 .). For all current predetermined regressors $x_{i,t}^{(m)}$ use the option gmm(1 .) and for all current exogenous regressors gmm(0 .), except for variables like the first difference of time dummies which each just yield one instrument namely the regressor itself. If this leads to too many instruments in relation to $N(T-1) - K$, where K is the total number of estimated coefficients, including the $T-1$ time effects, then use the collapse option for all variables, or

provided it is nonzero, any possible form of dynamics in a relationship can adequately been taken care of. That this is seriously short-sighted follows from Section 3.

curtail, by replacing for all gmm-statements the dot by 4, 3 or 2. Do this possibly jointly with collapsing. So, in principle, do not use fewer internal instruments obtained from exogenous[14] than from predetermined regressors, and not fewer from predetermined than from endogenous regressors. Well substantiated rules of thumb on reasonable boundaries on the number of instruments are not available, apart from the obvious restrictions that it should be somewhere between the number of estimated regression coefficients and the number of available observations $N(T-1)$. Denoting the total number of employed instruments by L, for the current candidate MSM L should (to my current subjective taste) obey (very roughly) inequalities like

$$K + 4 \leq L < q_K K \quad \text{and} \quad q_L L < NT - T - K, \tag{4.1}$$

where q_K and q_L may be in a wide range[15], possibly $4 < q_K < q_L < 10$. After estimation, do not yet verify and interpret the coefficient estimates seriously before stage 3 has been reached and an acceptable MSM may have been found. First, just verify for both 1-step and 2-step results whether they pass the requirements regarding the two tests on serial correlation and for the standard and all incremental Hansen[16] overidentification restrictions tests. As long as no evidence has been produced on homoskedasticity of the disturbances completely neglect the outcome of the Sargan test. The overall Hansen test, which has $L - K$ degrees of freedom, and the 1-step and 2-step based tests on the second-order serial correlation coefficient of the disturbances, should all have a p-value (well) above 0.20, say, and the 1-step and 2-step tests for the first-order serial correlation coefficient should have a p-value below, say, 0.05. Otherwise, reformulate the model and/or instrument set. Especially if any of the incremental Hansen tests has a p-value below 0.10, say, this may give a clue on how to reformulate the model by including extra lags of that variable as regressor in the model. Also consider changing the functional form of the relationship by employing variable transformations or include variables that had not been tried yet, or allow for nonconstancy of the impacts by including either dummy variables that define categories of individuals per time period or include squared regressors or interaction terms. This requires employing instruments from these dummies, squared variables and interactions from lag 2 onwards, or from lag 0 onwards if their exogeneity is self-evident. So, go back to stage 1, and reconsider how to obtain an improved candidate MSM which passes all the above criteria reasonably well. Especially because all our inference techniques are asymptotic in nature one should prefer the sample to be as large as possible. However, it could turn out to be easier to find a satisfactory model with constant coefficients by seeking a more homogeneous subsample, either by focussing just on a subgroup (for instance, only families with children, or just firms with more

[14] Many practitioners do just use one instrument per exogenous regressor, namely te regressor itself, whereas lags of these often constitute relatively strong instruments for other regressors.

[15] Underlying motivations for these inequalities are that L should be strictly larger than K to assure the existence of at least the first four moments of the GMM coefficient estimates. This may help to avoid the occurrence of huge outliers. Also, one may decide to use a relatively very large number of instruments, say $10K$, but only in situations where the number of degrees of freedom $N(T-1) - K$ is at least 10 times as large as L too.

[16] In Xtabond2 the Hansen (incremental) tests presented for robust 1-step GMM are in fact those calculated for 2-step GMM.

than 100 employees), so decreasing N. Or by focussing on a particular era (only the years after the financial crisis), so decreasing T. It might be the case that analyzing the complete available sample requires so many extra parameters that separate analysis of subpopulations proves beneficial.

3. If the above criteria have been fulfilled, and especially if none or very few of the incremental Hansen tests has a p-value below 0.1, say, whereas increasing $p_0, ..., p_M$ by one, or adding squares or interactions of variables, does not yield corresponding coefficients with p-values below 0.2, say, it could be that a reasonably adequate and acceptable MSM has been found. Next, one could move on to stage 4, or first verify whether any of the coefficients for the longest lag of a variable $x_{i,t}^{(m)}$ or of $y_{i,t}$ has a t-value below 0.5, say, or a p-value above 0.6 or 0.7, say. If so, impose the least significant one of them to be zero, re-estimate the model, and repeat the same procedure until the coefficients of all longest lags have absolute t-values (well) above 0.5, and the overall and incremental Hansen tests and the serial correlation tests still produce satisfactory results.

4. Optional: Obtain for both 1-step and 2-step results the level residuals and run regressions by LSDV (fixed effects or within least-squares estimation) for their squares on all levels of the regressors that remained after stage 3, except the unlagged ones that were treated as endogenous. If none of their coefficients has an absolute t-value above 1, say, this could mean that the disturbances are (almost) homoskedastic and that 1-step estimation without asking for robust standard errors for the model at the end of stage 3 may be taken serious too, provided the overall Sargan test has a p-value above 0.20, say, and also the incremental Hansen and serial correlation tests remain satisfactory. If accepting homoskedasticity seems reasonable then from now on just standard 1-step GMM could be employed.

5. Now examine sequentially (in case of homoskedasticity just for 1-step, otherwise just for corrected 2-step) one by one, for all unlagged explanatory variables yet classified as endogenous with respect to $\varepsilon_{i,t}$ what the p-value is of the incremental Hansen test for the instruments that would be valid too if that regressor were actually predetermined; this involves employing one by one the extra option gmm(1 1) for the instruments constructed from each variable initially treated as endogenous. For the variable with the highest such p-value, provided this value is above 0.5, say, one may contemplate treating it from now on as predetermined, hence adopting the extra instruments gmm(1 1). Next, repeat the same procedure until no such incremental p-values can be obtained with values above 0.5, say. During this process check whether the overall and other incremental Hansen tests and the serial correlation tests (which use coefficient standard errors, so it is essential that the Windmeijer correction has been applied) are still satisfactory.

6. Next, repeat stage 5, but now checking sequentially for all unlagged variables $x_{i,t}^{(m)}$ yet treated as predetermined with respect to $\varepsilon_{i,t}$ whether they actually seem exogenous; this involves employing the extra option gmm(0 0). After stages 5 and 6 due to now using possibly some extra valid and relatively strong instruments, ideally the standard errors of the coefficient estimates should be smaller now. If homoskedasticity has not been accepted then estimate the present specification

of model and instrument set not just by corrected 2-step but also by robust 1-step GMM. Hopefully the results are reasonably similar. If not, examine where in the process the results started to diverge. If a discrepancy cannot be resolved my subjective preference would be to continue on the basis of the robust 1-step results.

7. Now, again preferably sequentially, and as long as no problems emerge regarding the serial correlation and instrument validity tests, impose restrictions on the model by removing regressors with absolute t-ratio's below 0.5, say, starting with those with the highest p-values (smallest absolute t-values). Probably the effectivity of stages 5 and 6 would benefit when preceded by stage 7, but undoubtedly the power of the tests used in 7 will benefit after stages 5 and 6 yielded some extra strong and valid instruments. Anyhow, it could be beneficial to start stages 5 and 6 only after the most insignificant regressors have already been removed. Possibly it is worthwhile to examine any differences that occur when the examination of the acceptability of imposing zero coefficient restrictions and imposing extra orthogonality conditions is performed in different orders.

8. At this stage, but also depending on what seemed required already in stage 2, one should (aiming to reduce estimator bias) investigate the effects of restricting the number of employed instruments either by removing long lags (curtailing) or taking linear combinations of instruments (collapsing), or by both. In this process, keep an eye on the inequalities mentioned in stage 2, taking into account that K, L and T will/may be different since adopting the MSM in stage 3. Note that T here denotes the time-series sample size, being the number of available time-series observations for the left-hand side variable in the equation before taking first differences. This number may have increased since stage 3, due to imposing zero restrictions on the coefficients of the longest lags. And K will usually be smaller now and L larger.

9. Keep in mind that valid relevant coefficient restrictions are not necessarily just of the zero restriction type. Possibly, two coefficients may for good reasons have opposite sign, so sum to zero, or sum to unity, etc. Testing whether a total multiplier is zero may require to test the significance of the sum of a series of coefficients.[17] Compulsively removing regressors when their p-value exceeds 5% is condemnable. Demonstrating the insignificance of an effect can be very informative as such. Removing regressors with t-values (well) below 1 may make sense if there is no strong theory to leave them in. Useful additional evidence can be produced by also testing the joint significance of groups of single coefficient restrictions already imposed on the MSM and verifying whether the p-value is high indeed. Such joint significance tests can be obtained by using the "test" option.

10. Finally, or immediately after stage 7, a similar sequential procedure as in stages 6 and 7 can be performed regarding additional first-differenced instruments that under effect stationarity of an individual regressor $x_{i,t}^{(m)}$, which requires $E(\eta_i \Delta x_{i,t}^{(m)}) =$

[17] For instance, testing whether $\beta_0 + \beta_1 + \beta_2 = 0$ in the model $y_{it} = \gamma y_{i,t-1} + \beta_0 x_{it} + \beta_1 x_{i,t-1} + \beta_2 x_{i,t-2} + ...$ can be done by estimating $y_{it} = \gamma y_{i,t-1} + \beta_0^* \Delta x_{it} + \beta_1^* \Delta x_{i,t-1} + \beta_2^* x_{i,t-2} + ...$ and then testing the significance of β_2^*, because $\beta_0^* = \beta_0$, $\beta_1^* = \beta_1 + \beta_0$ and $\beta_2^* = \beta_2 + \beta_1 + \beta_0$.

0, would be valid for the model in levels. In the level equation Xtabond2 automatically includes an intercept term, both in the equation and in the set of instruments. Note that the (lagged) dependent variable $y_{i,t}$ can only be effect stationary if all the other regressors are effect stationary too while y_{i0} obeys the required initial conditions. Hence, instead of immediately testing the effect stationarity of all variables, one better starts by investigating the regressors $x_{i,t}^{(m)}$ one by one.[18] Only if all the associated p-values are, say, 0.3 or larger, it seems not unlikely that standard 2-step Blundell-Bond GMM (Windmeijer-corrected) should be preferred to an Arellano-Bond implementation. Mostly, Blundell-Bond estimates have smaller estimated standard errors than Arellano-Bond estimates anyhow, but this should not automatically be taken as evidence of their superiority; it could also be a side-effect of their inconsistency.[19] If not all regressors are found to be effect stationary it can nevertheless make sense to do system estimation (jointly estimating the equation in first-differences and the equation in levels), just exploiting the instruments in first-differenced form for the level equation obtained from those regressors $x_{i,t}^{(m)}$ which seem effect-stationary.

Of course, the above 10 stages should not be followed in a too mechanical way. Every particular data set and relationship goes with specific peculiarities that require special attention. Especially the p-value thresholds suggested in the above should be taken with wide margins. Choosing a really sensible value would require to first quantify the risks and costs of taking wrong decisions, and knowing the finite sample distribution of the test statistics (instead of just their asymptotic chi-squared null-distribution), which is practically impossible.

The model specification strategy formulated in the above 10 stages starts with finding after some trial and error an MSM for which no obvious evidence is found that its underlying statistical assumptions should be rejected, and next in a sequential search a restrained version of it is established in which coefficient restrictions have been imposed and extra moment conditions are employed to estimate it hopefully more accurately. From that final specification series of interim multipliers and any other substantive inferences regarding the established relationship should be obtained. One should realize, though, that it would be naive to interpret the ultimate findings on this restrained MSM in the usual standard way. Estimated standard errors will carry only little information now on the actual statistical variability of the estimated coefficients and a solid standard for the actual significance of the final results can not be provided. The final result is the product of data reduction instigated by reasonably sophisticated but at the same time rather cosmetic, say cosmetric, methods. They cannot simply lead to the construction of trustworthy 95% confidence intervals on the values of the unknown structural parameters. As discussed in Section 3, from the estimates of the restrained MSM short-run and long-run multipliers can be obtained, but a serious assessment of their standard errors seems an unattainable endeavour.

To conclude this section: Always keep in mind, that larger/smaller estimated standard errors between different implementations of an estimation technique do not provide

[18] How this is done is illustrated in de do-file given in Appendix B.

[19] For similar reasons OLS estimates usually have smaller estimated standard errors than IV estimates. This occurs irrespective whether OLS is consistent or not. To choose between them one should test the validity of the extra orthogonality conditions used by Blundell-Bond.

evidence on similar differences in actual accuracy. The difference in true standard deviation may be opposite and there may be a substantial difference in actual bias too. It usually occurs that when extra information is being exploited (coefficient restrictions imposed, moment conditions employed) this yields estimates with smaller estimated standard errors, irrespective whether the extra information is actually valid or not.

5. An empirical example

Following the principles and strategy outlined above we will illustrate these by reanalyzing the same sample of only $N = 140$ UK companies observed annually over a period not exceeding 1976-1984 that has earlier been examined in the classic studies by Arellano and Bond (1991) and Blundell and Bond (1998).[20] They considered a dynamic employment equation for the logarithm of employment ($n_{i,t}$) in company i at the end of year t. For its explanation the data set contains three variables ($M = 3$), namely the log of real product wage ($w_{i,t}$), the log of gross capital ($k_{i,t}$), and the log of industry output ($ys_{i,t}$), next to the linear trend ($year$).

Table 1 reports in its columns for particular model specifications the GMM coefficient estimates, with their standard errors between parentheses. The table does not present the time effects which were always included. The coefficient values have three stars when the absolute t-ratio is larger than 3, otherwise two stars when larger than 2, and otherwise one star when larger than 1. In the heading of each column codes are given indicating which particular implementation of GMM has been used, which classification of the three separate current regressors has been adopted, and whether any particular form of reduction regarding the available instruments has been used. Also the total number of regressors (K), the total number of instruments (L) and the asymptotic p-values of the major misspecification test statistics as calculated by Stata package Xtabond2 are given. If the data set would have been balanced over the period 1976-1984 there would have been $N(T-1) = 140*6 = 840$ observations for all columns in Table 1, which all concern specifications with second order dynamics. Because the actual data set is unbalanced only 611 observations of $\Delta n_{i,t}$ participate as left-hand side variable in estimation. For the code to obtain all presented and some extra supporting results see Appendix B.

Column (A) replicates the results from column (a1) of Table 4 in Arellano and Bond (1991, p.290). So, it gives 1-step Arellano-Bond estimates with heteroskedasticity robust standard errors (indicated as AB1R). The standard errors are slightly larger here due to applying a degrees of freedom correction. From the predetermined lagged dependent variable all 27 available instruments have been employed (indicated as Pa), whereas all three current regressors have been treated as exogenous, while these first differenced regressors and their lags have all directly been used as instruments, and thus have not been used to generate any extra instruments (XXXn, where the n indicates no over-identification). Due to the many instruments constructed from lags of the dependent variable the degree of over-identification is nevertheless 25=27+14-16.

[20] Other empirical examples inspired by the strategy outlined in the foregoing section can be found in Kiviet, Pleus and Poldermans (2017, p.43) and Kiviet, Pindado and Requejo (2019).

Table 1. Empirical findings for the Arellano-Bond (1991) data

	(A)	(B)	(C)	(D)	(E)	(I)	(J)
	AB1R	AB2	AB2W	AB1R	AB1R	AB1R	BB2W
	PaXXXn	PaXXXn	PaXXXn	PEEEa	PEEE4	PEEX32	PEEX32
$n_{i,t-1}$	$.686^{***}$	$.629^{***}$	$.629^{***}$	$.759^{***}$	$.913^{***}$	$.966^{***}$	1.07^{***}
	(.147)	(.092)	(.197)	(.082)	(.105)	(.098)	(.044)
$n_{i,t-2}$	$-.085^{*}$	$-.065^{**}$	$-.065^{*}$	$-.132^{**}$	$-.127^{**}$	$-.159^{***}$	$-.104^{**}$
	(.057)	(.027)	(.046)	(.046)	(.043)	(.049)	(.035)
$w_{i,t}$	$-.608^{***}$	$-.526^{***}$	$-.526^{***}$	$-.538^{***}$	$-.582^{***}$	$-.615^{***}$	$-.490^{***}$
	(.181)	(.055)	(.157)	(.158)	(.159)	(.114)	(.088)
$w_{i,t-1}$	$.393^{**}$	$.311^{***}$	$.311^{*}$	$.579^{***}$	$.721^{***}$	$.565^{***}$	$.519^{***}$
	(.171)	(.096)	(.206)	(.184)	(.226)	(.184)	(.107)
$w_{i,t-2}$	-	-	-	$-.100^{*}$	$-.161^{**}$	-9.52^{**}	-1.97^{*}
				(.065)	(.070)	(3.43)	(1.60)
$k_{i,t}$	$.357^{***}$	$.278^{***}$	$.278^{***}$	$.334^{***}$	$.239^{*}$	$.903^{*}$	0.993^{*}
	(.060)	(.046)	(.074)	(.100)	(.133)	(.692)	(.554)
$k_{i,t-1}$	-.058	.014	.014	$-.104^{*}$	$-.261^{**}$	-2.53^{**}	-1.87^{**}
	(.074)	(.054)	(.094)	(.081)	(.095)	(.866)	(.676)
$k_{i,t-2}$	-.020	$-.040^{*}$	-.040	-.019	-.030	1.46^{**}	$.718^{*}$
	(.033)	(.026)	(.044)	(.033)	(.037)	(.715)	(.428)
$ys_{i,t}$	$.609^{***}$	$.592^{***}$	$.592^{***}$	$.536^{**}$	$.895^{***}$	$.575^{**}$	$.383^{*}$
	(.175)	(.118)	(.176)	(.219)	(.277)	(.195)	(.201)
$ys_{i,t-1}$	$-.711^{***}$	$-.566^{***}$	$-.566^{**}$	$-.641^{**}$	$-.954^{***}$	$-.722^{***}$	$-.652^{**}$
	(.235)	(.142)	(.265)	(.223)	(.307)	(.238)	(.243)
$ys_{i,t-2}$	.106	.101	.101	$.230^{*}$	$.371^{*}$	-25.8^{*}	-17.2^{*}
	(.143)	(.114)	(.164)	(.180)	(.198)	(13.7)	(10.1)
$year$	.010	$.011^{*}$	.011	-	-	-	-
	(.010)	(.008)	(.012)				
K	16	16	16	17	17	25	26
L	41	41	41	114	74	100	173
$m1$	.000	.003	.034	.000	.000	.000	.000
$m2$	.606	.678	.725	.934	.531	.842	.799
$J^{(2,1)}$	.177	.177	.177	.532	.939	.927	.962

At least four aspects of (A) are surprising: (i) leaving out the $w_{i,t-2}$ regressor; (ii) adopting exogeneity for all three explanatories; (iii) using all possible instruments from $n_{i,t}$ and the absolute minimum from the other explanatories; (iv) including, next to the time dummies, a linear trend (which gives an intercept in the first-differenced model). Regarding (iv) the explanation is simple: if time-dummies had been included for all $T-1=6$ time-series observations in the estimation equation[21] then estimating the linear trend as well would have been problematic, but not when just 5 time dummies are included, as occurred. Regarding (iii) we would suggest using some further lags of the

[21] From the 9 observations during the 1976/1984 period one is lost due to taking first differences and another 2 are lost due to having twice lagged variables in the regression, so at most 6 remain. In this dataset missing observations do not occur halfway the sample period, but just at the beginning or the end.

three exogenous regressors as instruments too, because these will be useful explanatories of $\Delta n_{i,t-1}$ and $\Delta n_{i,t-2}$ in the implicit first-stage regressions. With respect to (ii), it seems that one better should not take the exogeneity of the three regressors for granted at this stage. Regarding (i), simply adding regressor $\Delta w_{i,t-2}$ to the first differenced model and employing it as extra instrument as well yields $-.146^{*}$ (.087) for its coefficient and standard error (this regression is not reported in the table, but is available via Appendix B), which makes one wonder why this regressor has been left out, and not the other three second order lags, which all have smaller absolute t-ratio's. The p-values for the autocorrelation tests $m1$ and $m2$ are satisfying, but the Hansen J test does not inspire great confidence.

In column (B) we replicate column (a2) from Table 4, which applies 2-step GMM (AB2). The results regarding the coefficients are again equivalent with those published earlier. In column (C) we estimate the same equation, but calculated the Windmeijer-corrected standard errors (AB2W), which were not yet developed in 1991. These are found to be often almost twice as large as the uncorrected ones. The robust 1-step and corrected 2-step results of columns (A) and (C) do not differ much. Extending the model of column (C) with regressor $w_{i,t-2}$ yields now $-.133^{*}$ (.086), so leaves the same doubts about its exclusion. The Hansen test, as calculated by Xtabond2, is exactly the same statistic in columns (A), (B) and (C). The serial correlation tests differ, because they depend on coefficient standard errors and residuals, but they are again satisfactory.

In columns (b), (c) and (d) of Table 4 in Arellano and Bond (1991, p.290) models with further restricted dynamics have been estimated by 2-step GMM upon treating in (c) and (d) w and k as endogenous now, whereas as instruments, next to their second lags, also some external instruments have been employed. Due to lack of detailed information these results are hard to replicate. From now on we will switch to the 10-stage strategy developed in Section 4.2 and examine the single relationship for $n_{i,t}$, just using internal instruments.

Column (D) of Table 1 contains 1-step robust estimates (AB1R) for the ADL(2,2,2,2) model, where all three explanatories have been treated as endogenous and all available internal instruments have been used (PEEEa). We removed the trend, which does not affect the slope coefficient estimates, because we always included $T-1=6$ time-dummies. Probably, employing 114 instruments for a total of only 611 observations, as in (D), may lead to detrimental finite sample problems due to overfitting the 11 non-exogenous regressors. Therefore we reestimate this model in column (E) using curtailing. We left out instruments beyond lag 4, which reduces their number by 40. Most standard errors in (E) are larger than in (D), as is to be expected, but also quite a few of the estimated coefficients differ between (D) and (E), those from (E) often about one standard error further away from zero. This might be due to (D) being more vulnerable to finite sample bias. In (E) the coefficient for the earlier omitted variable $w_{i,t-2}$ seems significant. The serial correlation tests are still satisfactory, whilst the Hansen test is much more comforting than before. In the model of column (E) we find for the incremental Hansen tests for the 17 instruments constructed from each of the variables n, w, k and ys the p-values 0.71, 0.72, 0.75 and 0.66 respectively. Hence, at first sight, it seems that we may have obtained here an adequate MSM.

Testing the restrictions which would simplify (E) to a simple partial adjustment model (removing all lags, apart from $n_{i,t-1}$) yields a p-value of 0.00. Despite the fact that partial adjustment is strongly rejected, estimating the partial adjustment model yields seemingly highly significant coefficients, whereas the p-values of $m1$ and $m2$ are 0.00 and 0.62, and 0.15 for the Hansen test. Hence, these three diagnostics do not appear very powerful to detect dynamic misspecification in comparison to the test for omitted regressors.

Comparing the implications of (A) and (E) on the basis of their multipliers we find that the impact multipliers (the coefficient estimates of $w_{i,t}$, $k_{i,t}$ and $ys_{i,t}$) do not differ very much. However, their total multipliers (3.2) vary a lot. For (A) TM_w^n, TM_k^n and TM_{ys}^n are -0.54, 0.25 and 0.01 respectively, whereas for (E) these are -0.10, -0.24 and 1.46. Note, though, that it is well-known that such ratio estimators do not have finite moments and may therefore vary widely, especially in relatively small samples. By testing hypotheses like $\Sigma_{l=0}^{p_m}\beta_l^{(m)} = 0$ we find that all three total multipliers for specification (E) are strongly significantly different from zero. Testing for ys the hypothesis $\Sigma_{l=0}^{p_m}\beta_l^{(m)} + \Sigma_{l=1}^{p0}\beta_l^{(0)} = 1$ yields p-value 0.63, so its total multiplier is not significantly different from 1.

To collect more evidence on the putative adequacy of model (E) we examine the p-values of tests for the significance of particular additional regressors. Adding to (E) four extra regressors, namely the four variables each lagged three times, reduces the number of observations from 611 to 471 and yields, when testing their joint significance, a p-value of 0.84. From this, and the satisfactory serial correlation tests in (E), we deduce that higher order lags than two do not seem required to model the dynamics.

But what about the functional form? We constructed the squares of the three variables w, k and ys, added them and their first and second lags to the specification of model (E) and also used them to construct instruments, treating them as endogenous. This increases the number of regressors to 17+9=26 and would increase the number of instruments to 74+3x17=125 which we suppose is too large. Therefore we take just second and third lagged variables as instruments (interacted with the time-dummies, as usual), which yields $L = 86$. Now the joint test of the extra 9 regressors has p-value 0.05, which falsifies (E) as representing a statistically adequate MSM, unless we decide to neglect this result because we suspect that it is due to unsatisfactory finite sample behavior.

We choose to accept that it seems well possible that no simple homogeneity assumptions on the dynamic effects of the three variables w, k and ys should be imposed. So, despite the fact that N is rather small here, we will embark on a more general analysis of the possible significance of interaction effects in this relationship. Therefore, also the variables $w * k$, $w * ys$ and $k * ys$, have been constructed. They will be treated, at least initially, as endogenous. In Table 2 we just present the coefficient estimates and standard errors of the (lagged) squared and interacted regressors. Including them with the same lags as the other explanatories leads to 26+9=35 regressors. If we just use second and third lags as instruments this leads to the pretty large total of 122 instruments. Using only the second order lags as instruments leads to just 64 instruments and much larger standard errors. When we collapse all the instruments in the standard way 76 instruments remain with results that differ substantially from those that simply skip higher-order lags from the full set of available instruments. Collapsing yields more insignificant regressors too. We also estimated the model with 35 coefficients where we used as instruments just lag two of the four original variables and both lags two and three for the squares and interactions. This leads to 100 instruments.

Table 2. Further empirical findings for the Arellano-Bond (1991) data

	(F)	(G)	(H)	(I)	(J)
	AB1R	AB1R	AB1R	AB1R	BB2W
	PEEE32	PEEE32	PEEX32	PEEX32	PEEX32
$w_{i,t} \times w_{i,t}$	-.170	-	-	-	-
	(.327)				
$w_{i,t-1} \times w_{i,t-1}$	.151	-	-	-	-
	(.298)				
$w_{i,t-2} \times w_{i,t-2}$	.372*	.156	.208*	.225*	-.199*
	(.289)	(.199)	(.176)	(.175)	(.100)
$w_{i,t} \times k_{i,t}$	.273***	.210**	.213***	.205***	.157***
	(.079)	(.073)	(.068)	(.068)	(.044)
$w_{i,t-1} \times k_{i,t-1}$	-.150**	-.074*	-.085*	-.080*	-.151***
	(.064)	(.064)	(.065)	(.064)	(.038)
$w_{i,t-2} \times k_{i,t-2}$	-.039	-	-	-	-
	(.086)				
$w_{i,t} \times ys_{i,t}$	-1.01*	-.265	-.061	-	-
	(.990)	(.521)	(.394)		
$w_{i,t-1} \times ys_{i,t-1}$	.953	-	-	-	-
	(1.01)				
$w_{i,t-2} \times ys_{i,t-2}$	1.21*	1.64**	1.73**	1.70**	.667*
	(.919)	(.676)	(.675)	(.660)	(.380)
$k_{i,t} \times k_{i,t}$	-.017	-	-	-	-
	(.022)				
$k_{i,t-1} \times k_{i,t-1}$	-.006	-	-	-	-
	(.023)				
$k_{i,t-2} \times k_{i,t-2}$	.016*	.010	.013	-	-
	(.014)	(.015)	(.015)		
$k_{i,t} \times ys_{i,t}$	-.185*	-.254**	-.227*	-.241*	-.238**
	(.145)	(.123)	(.120)	(.122)	(.104)
$k_{i,t-1} \times ys_{i,t-1}$	.574***	.588***	.572***	.562***	.450***
	(.179)	(.173)	(.173)	(.163)	(.133)
$k_{i,t-2} \times ys_{i,t-2}$	-.349*	-.378**	-.348**	-.321**	-.175*
	(.199)	(.178)	(.168)	(.153)	(.090)
$ys_{i,t} \times ys_{i,t}$	.295	-	-	-	-
	(1.30)				
$ys_{i,t-1} \times ys_{i,t-1}$	-2.21*	-1.96*	-0.023	-	-
	(1.72)	(1.41)	(.833)		
$ys_{i,t-2} \times ys_{i,t-2}$	3.90*	3.67*	2.36*	2.20*	1.64*
	(2.10)	(1.88)	(1.69)	(1.36)	(1.16)
K	35	28	28	25	26
L	88	88	100	100	173
$m1$	.000	.000	.000	.000	.000
$m2$	.678	.780	.789	.842	.799
$J^{(2,1)}$	.687	.599	.927	.927	.962

In column (F) we present the results for using both lags two and three for the four original variables and just lag two for the squares and interactions. This leads to 88 instruments and yields satisfactory p-values of the misspecification tests. Not without hesitation, we are inclined to accept this as our general adequate MSM. For seven of the squares and interactions in (F) we find an absolute t-ratio below 1. Testing the significance of these 7 regressors jointly yields p-value 0.91. In column (G) we present the results after imposing these restrictions. Specification (G) has three coefficients of squares and interactions with absolute t-value below 1. Testing them jointly yields p-value 0.63. Thus, we could decide to impose them, but we do not (yet) and first examine the classification of the three explanatories w, k and ys.

In their specifications (c) and (d) Arellano and Bond treated ys as exogenous, without testing. Exogeneity of ys in G would imply 12 extra instruments (not taking any consequences for the square of ys into account). Testing their validity yields a p-value of 0.999, so we accept, which yields the results in column H. Testing in separate stages now validity of the 6 extra instruments associated with possible predeterminedness of w and k yields p-values of 0.02 and 0.04 respectively. So, the endogeneity of w and k is endorsed.

Three of the coefficients of (H) associated with squares and interactions have absolute t-values below 1; two of them correspond with those having small absolute t-values in (G). Testing them jointly yields p-value 0.84, so we impose them. Now k^2 and its lags have been removed from (F), so we could decide to remove its instruments too, or keep them as external instruments. We chose the latter, which yields column (I). None of its coefficients has absolute t-value below 1. We performed an informal test for heteroskedasticity by running a fixed effects regression of the squared level residuals of (I) on first and second lags of n, w, k, ys and the time dummies and obtained several significant coefficients, so will not use 1-step AB without robustifying the standard errors. We also estimated the specification of (I) by 2-step Windmeijer-corrected GMM. This does not differ much from robust 1-step estimation. Hence, just following the guidance by statistical criteria, we could accept (I) as our final AB result.

The coefficient estimates for specification I of the (lags of) the variables n, w, k and ys are presented in Table 1. We note that some of these estimates, in particular for $w_{i,t-2}$ and $ys_{i,t-2}$, have rather exorbitant values. This already occurred from specification (F) onwards. This could of course just be an effect of the inclusion of the squared and interacted variables. We calculated the interim multipliers, adopting for each of the three explanatories time invariance at a level equal to the average over all available data points. This yields impact multipliers for (I), giving those for (F) between parentheses, of -0.73 (-0.82), -0.19 (-0.47) and 0.58 (4.11) for w, k and ys respectively and total multipliers equal to -2.48 (-2.54), -0.87 (-0.88) and -30.4 (-21.9). Whether these figures throw a new light on genuine economic phenomena, or are simply the result of estimators that may be consistent but just behave awkwardly in small samples, or that the available data and model selection procedure used are in fact incompatible, is difficult to say at this stage. Van den Doel and Kiviet (1994) provide evidence that estimates of long-run multipliers tend to converge to the same probability limit irrespective of the adequacy of the specification of the short-run dynamics by the model. This result, though, provides no comfort for interpreting the just mentioned estimates from a rather small sample.

Finally we examined whether system estimation and using differenced variables as instruments for the equation in levels seems appropriate here. In column (J) of Tables 1 and 2 the results are presented of applying standard Windmeijer-corrected 2-step estimation according to Blundell-Bond (BB2W) to the specification of (I), supplemented in the level equation with an intercept. As is usually the case, Blundell-Bond estimation yields smaller, often substantially smaller, standard errors. So, should we prefer result (J) over result (I)? Result (J) assumes all regressors to be effect stationary. The difference test for validity of the 70 associated instruments presented by Xtabond2 yields p-value 0.965. Should this induce acceptance of (J)? At first sight that seems legitimate. Nevertheless, we suppose that (J) should be rejected for the following reason. Variable n can only be effect stationary if all other regressors are effect stationary. However, if w, k or ys are effect-stationary (which requires that their correlation with η_i is time-invariant) then it seems highly unlikely that their squares and interactions will be too. Therefore we tested whether w, k and ys are individually effect-stationary. This yields p-values of 0.235, 0.003 and 0.180 respectively. From this we conclude that the standard form of Blundell-Bond estimation should better be avoided here.

In the last two columns of Table 4 in Blundell and Bond (1998) the very same data set has been examined. However, their specifications omit various of the regressors we have found to have low p-values in less restrained specifications. They completely leave out regressor ys, and also any second-order lags, squared variables and interactions have been omitted. However, they treat variables $w_{i,t}$ and $k_{i,t}$ both as endogenous, like we do. Due to the restrictions they impose, which are strongly rejected, our results falsify their simple specification. We can exactly replicate the AB1R estimates presented in their last but one column, but our BB2W findings differ substantially from those in their last column, despite using the same regressors and number of instruments (these results can be obtained via Appendix B too). By system estimation we find a long-run elasticity of employment with respect to capital in the restrained specification of $(.49 - .42)/(1 - .93) = 1$ which is substantially larger than the 0.79 reported by Blundell and Bond (1998, p.138), who also focus on primary multipliers (neglecting the endogeneity). They ratify their findings on the basis of satisfying values for the $m1$ and $m2$ tests and the overall Hansen overidentification test. For the latter test AB estimation yields a p-value of 0.21. However, for BB estimation they report results implying a Hansen p-value of just 0.13. This implies a p-value of 0.16 for the incremental Hansen test of validity of all the effect stationarity assumptions. These values seem rather weak supporting evidence for these very simple first-order dynamic models, certainly for BB estimation. Our calculations for the strongly restrained model yield 0.20 and 0.24 for these p-values, but testing validity of the 7 instruments implied by effect-stationarity of variable w only produces a p-value of 0.05. So, also for the strongly restrained model, we conclude that these data provide no support for applying standard Blundell-Bond estimation. Thus, at long last, our specification search process selects result (I) to represent our preferred characterization of the structural relationship for variable $n_{i,t}$, although it yields interim multipliers with rather awkward values.

6. Major conclusions

This paper aims to provide practitioners with a detailed structured framework for a great many issues that require attention when searching for an adequate model specification and GMM implementation for a single structural dynamic relationship on the basis of a microeconometric panel data set. It pays attention in particular to selecting the relevant regressors and any interactions between them, jointly with determining whether these regressors and their lags are correlated with the disturbances and possibly equicorrelated with the individual effects. From this it follows which instrumental variables obey credible orthogonality conditions, and can be exploited when producing statistical inference on the parameters that determine the dynamic multipliers of the relationship. For practitioners assessment of these multipliers is often the major research objective.

In the available panel data literature, both theoretical and empirical, it has usually been taken for granted that relationships are homogeneous in the slope coefficients, at least over subsamples, whereas very little attention has been paid to the relatively simple possibilities offered by interaction terms to detect and eventually model slope heterogeneity. Recently attention has been given by theorists in particular to attempts to repair poorly specified panel models, not by adding constructive actually observed explanatories like interactions, but by supplementing the unobserved individual and time effects by further unobserved stochastic factors. Of course, only when these constructed factors sufficiently represent the wrongly omitted explanatories, including their relationship with the included regressors, this may yield a satisfying cure for misspecified models.

The extensive literature on variable selection strategies for regression analysis is of little help in the present dynamic panel data context, because this literature is mainly about situations where all candidate regressors are assumed to be exogenous so that application of GMM and selecting appropriate instrumental variables at the same time is not incorporated in those strategies. For the very same reasons we do not believe that automated search machines can easily be designed for the present multilayered problem.

A striking result from the empirical illustration is that in all the different formulations for the relationship estimated, also those which patently suffer from the omission of strongly significant regressors, the serial correlation tests $m1$ and $m2$ always gave satisfactory results. And also the overall J test of the over-identification restrictions demonstrates its meagre power to detect misspecified models when the number of instruments is substantially larger than the number of estimated coefficients. Only when difference Hansen tests are being used to test the validity of a limited subset of the instruments rejections have been obtained. Such rejections, assuming that they do not establish type I errors, are indispensable for steering the model selection process away from improper models, which either omit important explanatories or adopt invalid orthogonality conditions.

Both the empirical results for the classical data set examined in this paper and the experimental results obtained from the extensive simulation study in Kiviet et al. (2017) suggest that the present understanding of the actual properties of the current techniques for statistical inference on microeconomic panel data sets require an actual sample size in the cross-section dimension which is much larger than often available and used in practice. To uplift the accuracy of panel data inference it seems required that either much larger samples are being analyzed or that more sophisticated techniques are being developed. Any model selection strategy will fail when the null distribution of test statistics deviates substantially from their asymptotic approximations and coefficient estimates and the estimates of their standard deviation are seriously biased.

Acknowledgements

This paper benefitted greatly by having had the privilege to teach for many years Ph.D. courses in micro panel data analysis for graduate students at various places (Amsterdam, Innsbrück, Lisbon, Salamanca, Singapore, Stellenbosch, Zürich) and is also the result of many much appreciated discussions and interactions with a great number of colleagues and former students. In alphabetical order I would like to thank in particular: Maurice Bun, Julio Pindado, Milan Pleus, Rutger Poldermans, Ignacio Puerto and Frank Windmeijer. Of course, I bear sole responsibility for the contents of this study.

References

Arellano, M., Bond, S.R., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies* 58, 277-297.

Blundell, R., Bond, S., 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87, 115-143.

Bond, S.R., 2002. Dynamic panel data models: A guide to micro data methods and practice. *Portuguese Economic Journal* 1, 141-162.

Bun, M.J.G, Kiviet, J.F., 2006. The effects of dynamic feedbacks on LS and MM estimator accuracy in panel data models. *Journal of Econometrics* 132, 409-444.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31, 337-350.

Harvey, A.C., 1990. *The Econometric Analysis of Time Series, second edition.* Philip Allan, New York.

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

Hendry, D.F., 1995. *Dynamic Econometrics.* Oxford University Press, Oxford, UK.

Kiviet, J.F., 2017. Discriminating between (in)valid external instruments and (in)valid exclusion restrictions. *Journal of Econometric Methods* 6(1), 1-9.

Kiviet, J.F., 2019. Testing the impossible: identifying exclusion restrictions. Forthcoming in *The Journal of Econometrics.*

Kiviet, J.F., Phillips, G.D.A., 1986. Testing strategies for model specification. *Applied Mathematics and Computation* 20, 237-269.

Kiviet, J.F., Pindado, J., Requejo, I., 2019. Specification of dynamic panel data models: An empirical application to corporate finance. Draft mimeo.

Kiviet, J.F., Pleus, M., Poldermans, R.W., 2017. Accuracy and efficiency of various GMM inference techniques in dynamic micro panel data models. *Econometrics* 5(1), 14.

Roodman, D., 2009a. How to do xtabond2: An introduction to difference and system GMM in Stata. *The Stata Journal* 9, 86-136.

Roodman, D., 2009b. A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics* 71, 135-158.

Spanos, A., 2018. Mis-specification testing in retrospect. *Journal of Economic Surveys* 32, 541-577.

Van den Doel, I.T., Kiviet, J.F., 1994. Asymptotic consequences of neglected dynamics in individual effects models. *Statistica Neerlandica* 48, 71-85.

Wasserstein, R.L., Lazar, N.A., 2016. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* 70, 129-133.

Windmeijer, F., 2005. A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics* 126, 25-51.

Appendices

A. More on asymptotic p-values

Here we discuss and illustrate the difficulties faced when interpreting an individual p-value obtained from a single asymptotic test, and also on using series of p-values in the model selection strategy developed above. As is usually the case in econometrics, the limiting distribution under the tested null hypothesis of a test statistic is χ^2_{df} with a number of degrees of freedom (df) equal to the number of tested restrictions. When just 1 restriction is being tested this distribution specializes to χ^2_1, the square of a standard normal distribution. In that case it is possible to test the null hypothesis against one sided alternatives. Below we will just focus on the case of testing against two sided alternatives for $df \geq 1$. For some particular df the density function of χ^2_{df} is sketched in Figure 1 (green uninterrupted line). The figure shows two more densities, also just defined for positive scalar arguments; self-evidently all three density functions integrate to unity.

The figure indicates the argument $\chi^2_{(df,\alpha)}$; this is the critical value to be used for chosen nominal probability α of type I errors (rejecting the restrictions when they are true), where $0 < \alpha < 1$. The surface below the green density to the right of the α-level critical value $\chi^2_{(df,\alpha)}$ equals α. It is also called the nominal significance level of the test. In finite samples the actual distribution of a test statistic assuming the null hypothesis is true will as a rule differ from its asymptotic approximation. Its distribution is also determined by the unrestricted and generally unknown parameters of the data generating process. The discrepancy from the asymptotic null distribution depends on the values of the so-called nuisance parameters, but this dependence gradually vanishes the larger the sample is. The actual distribution of the test statistic under validity of the restrictions specified by the null hypothesis is represented in the figure by the red (dashed) density. The actual distribution of the test statistic for the true data generating process (without restricting it by the null hypothesis) is represented by the blue (dotted) density. The latter two densities are in practice generally unknown (but not necessarily in controlled computer simulations and in oversimplified text book models). Because the red and blue densities do not coincide in the figure we apparently consider a case where the null hypothesis is not true.

When we apply the test statistic we draw one realization from the blue (dotted) density and we compare it either with critical value $\chi^2_{(df,\alpha)}$, or calculate its so-called p-value and compare it with some threshold value p_c. Equivalence occurs when choosing $p_c = \alpha$. In case the null hypothesis is true then asymptotically (and thus hypothetically) the probability to draw a value to the right of $\chi^2_{(df,\alpha)}$ is α. In our actual sample this probability equals the surface under the red (dashed) line to the right of $\chi^2_{(df,\alpha)}$. The figure depicts a case where the test is apparently oversized, because the actual significance level is larger than the nominal significance level. Because in practice the actual null distribution is usually unknown we do not know whether an asymptotic test is undersized or oversized and neither do we know how much the actual and nominal significance level actually differ. Sometimes alternative asymptotically equivalent versions of test statistics are available which attempt to mitigate these size problems, either by simple degrees of freedom corrections or more sophisticated bias corrections, or by bootstrapping the test.

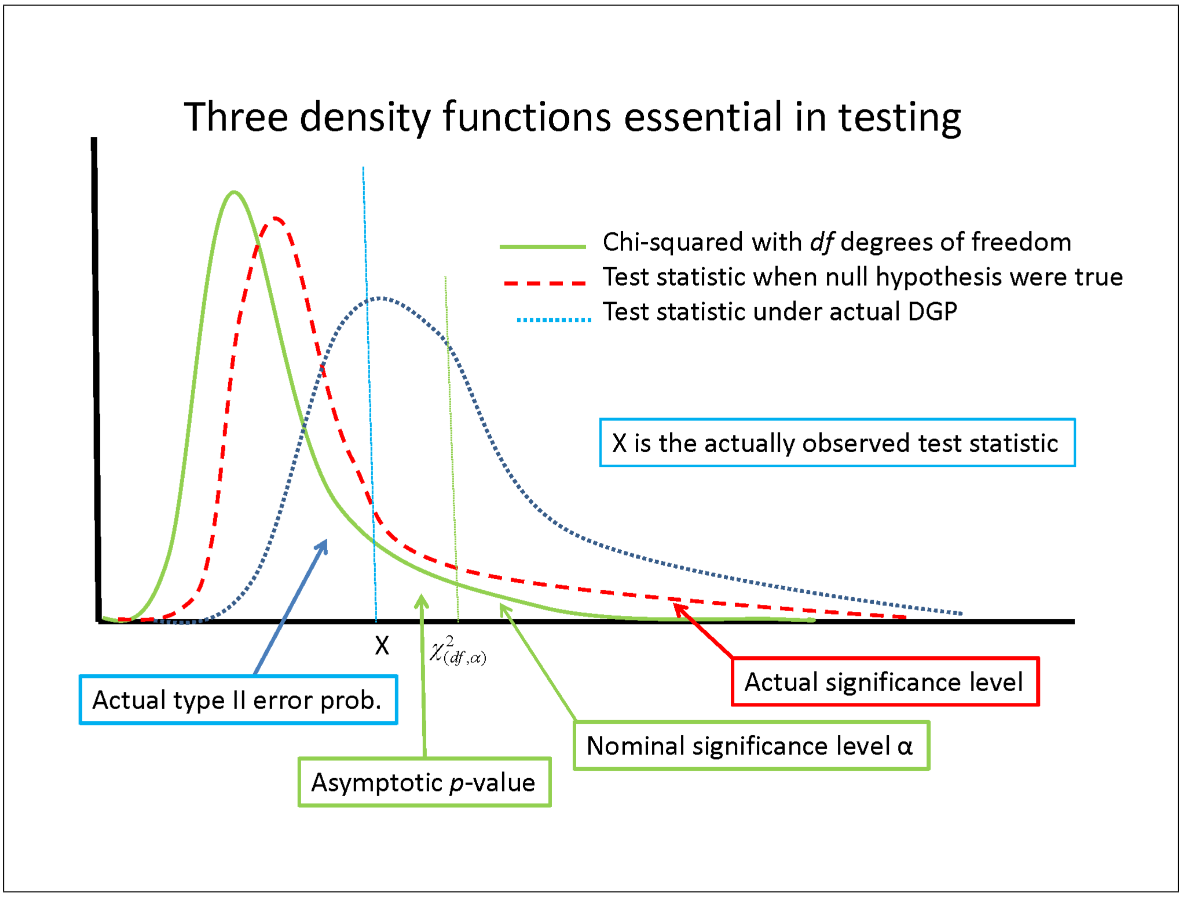
Three density functions essential in testing
Chi-squared with *df* degrees of freedom
Test statistic when null hypothesis were true
Test statistic under actual DGP
X is the actually observed test statistic
X
$\chi^2_{(df,\alpha)}$
Actual significance level
Actual type II error prob.
Nominal significance level α
Asymptotic *p*-value

Figure 1: Illustration of asymptotic tests, p-values, and significance.

Let us now assume that our single draw from the distribution of the test statistic (hence from the blue-dotted one) has realization indicated by X. The asymptotic p-value (as calculated by software, realizing $0 \leq p \leq 1$) is then obtained as the surface under the green density to the right of the observed test statistic value X. The unknown true p-value should be obtained by calculating the surface to right of X under the red (dashed) density, but this density being unavailable in practice one uses the asymptotic p-value, not knowing how (in)accurate it actually is.

Also unknown is the actual type II error (not rejecting a false hypothesis) probability of the test. It is the surface under the blue (dotted) density to the left of the available critical value $\chi^2_{(df,\alpha)}$. Note that this probability will be small if the blue (dotted) density is located much more to the right (which in principle corresponds to testing restrictions on parameter values which are very far from the truth). We cannot influence the unknown blue density, but we can choose a different critical value. By increasing α we move the critical value to the left and realize a smaller type II error probability at the expense of allowing the type I error probability to be larger. This is why we did plead in the foregoing for using much larger α values than, say, 0.05 in cases where committing a type I error has much less serious consequences than committing a type II error.

The power of a test is only formally defined for cases where the size of the test, which is the actual rejection probability of the test under the null hypothesis when maximized over all possible nuisance parameter values, can be controlled. If this is possible the test is called exact. If the red (dashed) density would be known true α-level critical values could be constructed.

In all stages of the model selection strategy developed in this study a great number of diagnostic, classification and specification tests are being used and all the time their asymptotic p-value is being used as a beacon to determine the direction in which the specification search process should proceed. Using p-values has the advantage that one does not have to look up critical values of the asymptotic null distribution of the test all the time.

B. Stata and Xtabond2 code and log files of the illustration

By the Xtabond2 package for Stata balanced and unbalanced panel data sets can be analyzed by estimating and testing linear models such as (2.1) and (3.17) using various implementations of GMM. See Roodman (2009a) for further detailed instructions on the various commands and their options. To actually use the package Xtabond2 it has to be installed as follows. While running Stata, give the command "ssc install xtabond2". A series of Stata and Xtabond2 commands can be collected in a so-called Stata do-file, and the output on the resulting statistical analysis can be collected in a so-called log-file.

The do-file ABexample.do, which generates all the estimation and test results mentioned in Section 5 and some useful additional ones, and its log-file ABexample.log, which contains all these results, can be downloaded from the authors personal homepage at: https://sites.google.com/site/homepagejfk/discussion-papers