The Stata Journal (2018)
18, Number 1, pp. 76–100

Testing for serial correlation in fixed-effects panel models

Jesse Wursten
Faculty of Economics and Business
KU Leuven
Leuven, Belgium
jesse.wursten@kuleuven.be

Abstract. Current serial correlation tests for panel models are cumbersome to use, not suited for fixed-effects models, or limited to first-order autocorrelation. To fill this gap, I implement three recently developed tests.

Keywords: st0514, xtqptest, xthrtest, xtistest, serial correlation, panel time series, fixed effects, higher-order serial correlation

1 Introduction

The issue of serial correlation in panel models has been largely ignored in recent decades (Inoue and Solon 2006), and the discussion is generally dismissed by using robust or clustered standard errors. However, as shown in Pesaran and Smith (1995), serial correlation can lead to inconsistent estimates in dynamic panels, and omitting or including time trends can dramatically change parameter estimates (for example, Allegretto, Dube, and Reich [2010]; Solon [1984]). Serial correlation tests can help to identify which model is most credible from a statistical perspective to complement arguments based on theory.

The two main commands currently implemented in Stata, `xtserial` and `abar`, are limited in their usability. `xtserial` implements the Wooldridge–Drukker (WD) (Drukker 2003; Wooldridge 2010) test, which is limited to first-order autocorrelation[1] and assumes a constant variance over time. Additionally, it is cumbersome to use when working with many or large models. I show in section 6 that the tests described in this article have considerably higher power than the WD test using Monte Carlo simulations.

The `abar` command, developed by Arellano and Bond (1991) and implemented in Stata by Roodman (2009), can test for any order of serial correlation but is not appropriate for fixed-effects regressions. In its current implementation, `abar` cannot be used after `xtreg`.

Instead, I propose using tests developed by Inoue and Solon (2006) and Born and Breitung (2016), which I have implemented as `xtistest`, `xtqptest`, and `xthrtest`. They are residual-based tests, which makes them fast and easy to use. Moreover, both `xtistest` and `xtqptest` can test for serial correlation up to any order. `xthrtest` is suited to detect first-order correlation only but relaxes the constant variance assumption.

1. More precisely, it can detect differences only between first- and second-order correlation.

By making serial correlation testing more accessible, I hope it will be used more frequently to test model specifications and refine the empirical framework. Often, the barrier of manually implementing such tests prevents theoretical advances from being used in applied research (Pullenayegum et al. 2016; Choodari-Oskooei and Morris 2016).

The remainder of this article is structured as follows. Section 2 describes the econometrics behind the tests, focusing on what they do and how they work. Section 3 discusses the syntax of the implemented commands. Section 4 demonstrates each test with a worked example. Section 5 provides some short guidelines on when to use which test. The relative strengths and shortcomings are detailed further using Monte Carlo evidence in section 6, where we also investigate the tests' performance when their assumptions are not met. Section 7 concludes.

2 Econometrics

2.1 xtistest: Inoue–Solon test for serial correlation

`xtistest` implements the portmanteau test for serial correlation introduced by Inoue and Solon (2006). More specifically, it tests whether any off-diagonal element of the autocovariance matrix $E(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_i')$ is nonzero, where $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{iT}]'$ is a vector of error terms. Thus, the unrestricted alternative hypothesis (H_a) is the presence of "some" serial correlation, without restricting the order at which it might occur, much like the ubiquitous Ljung–Box test used in time series. The drawback of this method is that as the number of elements in the covariance matrix increases quadratically with the length of the panel (T), so does the number of implicit hypotheses tested. Therefore, the test requires the number of panel units (N) to be large relative to the number of time periods.

Figure 1, panel A illustrates this graphically for a $T = 5$ panel, where the axes represent the time period and the lines represent the corresponding autocovariances tested. Different patterns indicate the covariances are used to test a different implicit hypothesis. In the current case, each covariance corresponds to a separate hypothesis. In total, there are $(T-1)\times(T-2)/2 = 6$ hypotheses.[2] Under the null hypothesis of no serial correlation, the IS statistic converges to a χ^2 distribution with $(T-1)\times(T-2)/2 = 6$ degrees of freedom as N goes to infinity. We refer to the original article for an exposition on how the test statistic is constructed concretely.[3]

2. Because of the fixed-effects framework, the covariances have to be demeaned first, which leads to a singular covariance matrix. To overcome this problem, Inoue and Solon (2006) drop one row and column from the covariance matrix. In the visual example (and in the code), I drop the last column to retain as much data as possible. The choice of column does not affect the test statistic ex ante.
3. Alternatively, Born and Breitung (2016) provide a more applied explanation.

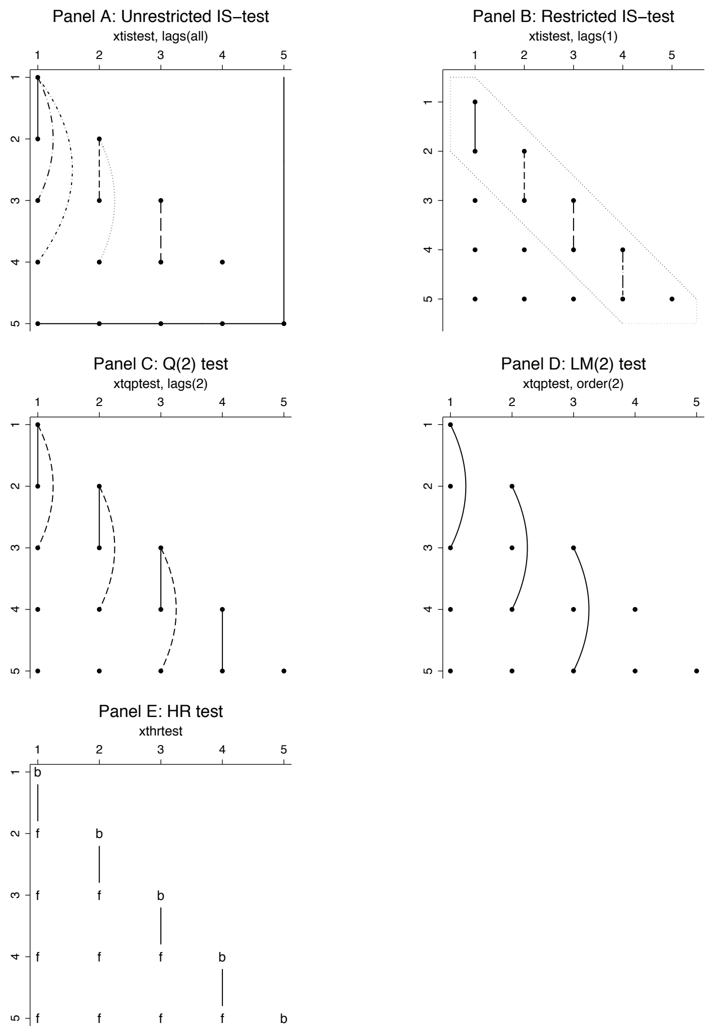

Figure 1. Visual representation of the various serial correlation tests. Each line represents the demeaned covariance between residuals at two different time periods. For example, a line between $(1,1)$ and $(1,3)$ refers to the covariance between $t = 1$ and $t = 3$. Likewise, $(3,3)$ to $(3,4)$ refers to the covariance between residuals at time $t = 3$ and $t = 4$. Different line patterns indicate that the covariances are used to test a different underlying hypothesis. The b's and f's relate to backward- and forward-demeaned residuals (see section 2.3).

As in the Ljung–Box test, the lag lengths considered in the IS test can be restricted to keep the number of hypotheses in check. Under the restricted H_0 of no autocorrelation up to lag p, the modified IS(p)-statistic is $\chi^2\{pT - p(p+1)/2\}$ distributed, growing only linearly in T. This is visually illustrated in panel B of figure 1 for $p = 1$. Only autocovariances of the first order are considered, leading to four underlying hypotheses to evaluate.

2.2 xtqptest

The `xtqptest` command calculates the two bias-corrected test statistics introduced by Born and Breitung (2016), referred to as LM(k) and $Q(p)$. The first tests for autocorrelation of order k, whereas the second looks for autocorrelation up to order p.

Lagrange multiplier (LM) test of serial correlation at order k

The LM(k) statistic is fairly straightforward and comes down to a heteroskedasticity- and autocorrelation-robust t test of $\varsigma = -1/(T-1)$, with ς as the coefficient on the kth-order demeaned residuals in (1). The e_{it} residual includes the fixed effects; that is, it is produced in Stata by `predict, ue`.

$$e_{it} - \overline{e}_i = \varsigma(e_{i,t-k} - \overline{e}_i) + \epsilon_{it} \tag{1}$$

This leads to the asymptotically equivalent test statistic $\widetilde{\text{LM}}(k)$ defined by (2) and (3), which under the null hypothesis of no serial correlation at order k has a standard normal limiting distribution and is calculated by `xtqptest, order(`k`)`.

$$z_{k,i} = \sum_{t=k+1}^{T} \left\{ (e_{it} - \overline{e}_i)(e_{i,t-k} - \overline{e}_i) + \frac{1}{T-1}(e_{i,t-k} - \overline{e}_i)^2 \right\} \tag{2}$$

$$\widetilde{\text{LM}}_k = \frac{\sum_{i=1}^{N} z_{k,i}}{\sqrt{\sum_{i=1}^{N} z_{k,i}^2 - \frac{1}{N}\left(\sum_{i=1}^{N} z_{k,i}\right)^2}} \tag{3}$$

Figure 1, panel D illustrates this visually for $k = 2$. There is now just one underlying hypothesis as the information from all autocovariances of lag length 2 is pooled. Note the contrast with the IS test, where they were considered separately.

Q test of serial correlation up to order p

The $Q(p)$ statistic, which tests for autocorrelation up to order p is more complicated because of finite-sample bias induced by correlation among the variously lagged demeaned residuals. Born and Breitung (2016) eliminate this bias by transforming the residuals as per (4).

$$L_k\mathbf{e}_i = \begin{pmatrix} \mathbf{0}_k \\ e_{i1} - \overline{e}_i \\ \vdots \\ e_{i,T-k} - \overline{e}_i \end{pmatrix} + \left(\frac{T-k}{T^2-T}\right)\mathbf{e}_i \tag{4}$$

The first vector on the right-hand side represents the lagged demeaned residuals ($\mathbf{0}_k$ is a column vector of k zeros). The second term, which involves the residuals including the fixed effects, removes the bias. It converges to zero as T increases and hence mainly plays a role in shorter panels. Ordinary least-squares estimates of the coefficients in (5) converge to zero for all T and $N \to \infty$ under the null hypothesis of no serial correlation.

$$\begin{aligned} \mathbf{M} &= \mathbf{I}_T - \mathbf{i}_T\mathbf{i}_T' \\ \mathbf{M}\mathbf{e}_i &= \phi_1 L_1\mathbf{e}_i + \phi_2 L_2\mathbf{e}_i + \cdots + \phi_p L_p\mathbf{e}_i + \mathbf{v}_i \end{aligned} \tag{5}$$

The matrix $\mathbf{M}$ demeans the fixed-effect residuals ($\mathbf{I}_T$ is the identity matrix, $\mathbf{i}_T$ a column vector of ones). Figure 1, panel C visualizes this test for p set to 2. As observations are pooled per lag length, there are two implicit hypotheses to test, $\phi_1 = 0$ and $\phi_2 = 0$. The `xtqptest, lags(`p`)` command calculates the asymptotically equivalent $\widetilde{Q}(p)$ statistic, defined in (6). Under the null of no serial correlation up to order p, it follows a χ^2 distribution with p degrees of freedom.

$$\begin{aligned} \mathbf{Z}_i &= (L_1\mathbf{e}_i, \ldots, L_p\mathbf{e}_i) \\ \mathbf{A}_i &= \mathbf{e}_i'\mathbf{M}\mathbf{Z}_i \\ \widetilde{\mathbf{Q}}_p &= \left(\sum_{i=1}^N \mathbf{A}_i\right) \left\{ \left(\sum_{i=1}^N \mathbf{A}_i'\mathbf{A}_i\right) - \frac{1}{N}\left(\sum_{i=1}^N \mathbf{A}_i'\right)\left(\sum_{i=1}^N \mathbf{A}_i\right) \right\}^{-1} \left(\sum_{i=1}^N \mathbf{A}_i'\right) \end{aligned} \tag{6}$$

2.3 xthrtest

`xthrtest` implements the heteroskedasticity-robust test statistic, also introduced in Born and Breitung (2016). It is based, respectively, on the forward- and backward-transformed residuals e_{it}^f and e_{it}^b defined in (7) and (8).

$$e_{it}^f = e_{it} - \frac{1}{T-t+1}(e_{it} + \cdots + e_{iT}) \tag{7}$$

$$e_{it}^b = e_{it} - \frac{1}{t}(e_{i1} + \cdots + e_{it}) \tag{8}$$

The statistic can then be seen as a heteroskedasticity- and autocorrelation-robust t test on the ψ coefficient in (9).

$$e_{it}^f = \psi e_{i,t-1}^b + \omega_{it} \qquad t \in \{3, 4, \ldots, T-1\} \tag{9}$$

Subtracting, respectively, the forward- and backward-looking mean in (7) and (8) ensures the test is robust to time-varying variances. Under the null of no first-order autocorrelation, it has a standard normal limiting distribution. Figure 1, panel E presents a visual representation of the heteroskedasticity-robust (HR) test. The dots have been replaced by b's and f's to highlight that the statistic is not based on the same demeaning process as the other tests.

3 Syntax

`xtistest` [*varlist*] [*if*] [*in*] [, `lags(`*integer*|`all)` `original`]

`xtqptest` [*varlist*] [*if*] [*in*] [, `lags(`*integer*`)` `order(`*integer*`)` `force`]

`xthrtest` [*varlist*] [*if*] [*in*] [, `force`]

The three commands share a similar syntax, which calculates their respective test statistics and p-values for the variables specified in *varlist*. Alternatively, they can be used as `xtreg` postestimation commands by omitting the *varlist*.

`xtistest` has two options. The first, `lags()`, indicates the maximum number of lags to check for autocorrelation. All lags up to this maximum are checked. Specifying `lags(all)`, leads to the unrestricted IS test, which checks for serial correlation of any order (see section 2.1). The test defaults to a maximum lag length of two (`lags(2)`).[4] By default, the command uses the faster Born and Breitung (2016) implementation, but the initial Inoue and Solon (2006) method can be requested by using the `original` option. The results should be identical.

`xtqptest` takes the mutually exclusive `lags()` and `order()` options. The first will calculate the $Q(p)$ test statistic for serial correlation up to order p, whereas the latter refers to the LM(k) test for serial correlation of order k. For example, if there is first- but not second-order autocorrelation, the test should reject the null if `lags(2)` was specified but not when the user opted for `order(2)`. The test requires *varlist* to be residuals including the fixed effect (`‘ue’` residuals in `predict` terminology). There is some internal machinery to check whether you actually supplied `‘ue’` residuals, but it is not infallible (in both directions). Specifying `force` skips this check.[5]

`xthrtest` can detect only first-order serial correlation and thus does not ask for any lag specification. The `force` option again bypasses the check described above.

4. The choice for lag length two is arbitrary. The alternative—making the unrestricted test the default—is unattractive because in very large datasets, calculating this statistic can take a significant amount of time.
5. Note that as T increases, the difference between using the standard residuals and the `‘ue’` residuals converges to zero.

4 Example

This section serves two purposes. First, it shows how the commands work in practice and how to interpret their output. Second, it highlights that coefficients can vary strongly depending on the modeling assumptions. Serial correlation measures by themselves cannot tell us which estimates are more credible; nonetheless, they are one of the elements that can guide this decision. For example, ceteris paribus, one would put more stock in a "correctly" specified model than one that exhibits clear statistical issues.

The example is based on a publicly available dataset of UK firms' yearly employment, wages, capital, and output between 1976 and 1984.[6] The same data were used in Arellano and Bond (1991) to illustrate the famous Arellano–Bond estimator and introduce their serial correlation test, implemented in Stata as `abar` (see section 1). As a first step, we regress the level of log employment (`n`) on the logs of wages (`w`), capital (`k`), and industry output (`ys`). We include year and firm fixed effects.

```
. webuse abdata

. xtreg n w k ys yr*, vce(cluster id) fe
note: yr1984 omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =      1,031
Group variable: id                              Number of groups  =        140

R-sq:                                           Obs per group:
     within  = 0.6321                                         min =          7
     between = 0.8481                                         avg =        7.4
     overall = 0.8349                                         max =          9

                                                F(11,139)         =      48.94
corr(u_i, Xb)  = 0.5930                         Prob > F          =     0.0000

                                  (Std. Err. adjusted for 140 clusters in id)
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |  -.2968768   .1262997    -2.35   0.020    -.5465938   -.0471597
           k |   .5475598    .050709    10.80   0.000     .4472991    .6478204
          ys |   .2648254   .1529614     1.73   0.086    -.0376065    .5672574
  (output omitted)
```

We find elasticities of -0.3, 0.55, and 0.26 for earnings, capital, and industrial output, respectively. However, our simple equation in levels might be misspecified. Perhaps there are strong trends in firm employment. If these are correlated to any of our independent variables, this would bias their coefficient estimates. One way to check whether this could be an issue is to look for serial correlation in the residuals. Let's start with first-order auto correlation, using the $Q(1)$ test.

6. The dataset is loaded in the first line of the example.

```
. xtqptest, lags(1)
Bias-corrected Born and Breitung (2016) Q(p)-test as postestimation
Panelvar: id
Timevar: year
p (lags): 1
```

Variable	Q(p)-stat	p-value	N	maxT	balance?
Post Estimation	65.17	0.000	140	9	unbalanced

```
Notes: Under H0, Q(p) ~ chi2(p)
   H0: No serial correlation up to order p.
   Ha: Some serial correlation up to order p.
```

The output of `xtqptest, lags(1)` lists the specific test performed, the structure of your panel, and the number of lags used ($p = 1$). Because no *varlist* was specified, the command predicted the fixed-effect residuals for us, based on the last regression. The test strongly rejects the null hypothesis, indicating the residuals are serially correlated. Next, we look at the output from the other tests [LM(1), IS(1), and HR].

```
. xtqptest, order(1)
Bias-corrected Born and Breitung (2016) LM(k)-test as postestimation
Panelvar: id
Timevar: year
k (order): 1
```

Variable	LM(k)-stat	p-value	N	maxT	balance?
Post Estimation	8.05	0.000	140	9	unbalanced

```
Notes: Under H0, LM(k) ~ N(0,1)
   H0: No serial correlation of order k.
   Ha: Some serial correlation of order k.
. xtistest, lags(1)
Inoue and Solo (2006) IS-test as postestimation
Panelvar: id
Timevar: year
p (lags): 1
```

Variable	IS-stat	p-value	N	maxT	balance?
Post Estimation	62.08	0.000	140	9	unbalanced

```
Notes: Under H0, IS ~ chi2(p*T-p(p+1)/2)
   H0: No auto-correlation up to order 1.
   Ha: Auto-correlation up to order 1.
. xthrtest
Heteroskedasticity-robust Born and Breitung (2016) HR-test as postestimation
Panelvar: id
Timevar: year
```

Variable	HR-stat	p-value	N	maxT	balance?
Post Estimation	1.31	0.190	140	9	unbalanced

```
Notes: Under H0, HR ~ N(0,1)
   H0: No first-order serial correlation.
   Ha: Some first-order serial correlation.
```

The output is organized in the same way as before. Both the LM and IS tests strongly suggest first-order serial correlation is present in the residuals. Rather surprisingly, the HR test suggests the residuals are not correlated at the first order. If there are good reasons to assume the variance changed significantly over the sample, then a conflicting result from the HR test should not be discarded easily.[7] In this case, however, there is no reason to assume the variance would change over time. Additionally, the other tests firmly rejected the null, with very extreme statistics. Combined, this makes the presence of first-order serial correlation more plausible than its absence.

The remainder of this section is summarized in table 1, where I present coefficient estimates for various specifications and their respective serial correlation results.[8] The numbers in parentheses represent the p-values corresponding to the coefficient estimates and serial correlation tests. Column (1) recapitulates the discussion above.

Table 1. Summary of various specifications

	(1) Levels	(2) Trends	(3) Differences	(4) Differences with Lags
Coefficient estimates				
w	−0.30 (0.02)	−0.39 (0.02)	−0.49 (0.00)	−0.43 (0.11)
k	0.55 (0.00)	0.41 (0.00)	0.33 (0.00)	0.52 (0.00)
ys	0.26 (0.09)	0.43 (0.09)	0.60 (0.00)	0.52 (0.14)
Serial correlation tests				
$Q(1)$	65.17 (0.00)	13.57 (0.00)	4.85 (0.03)	0.39 (0.53)
LM(1)	8.05 (0.00)	3.73 (0.00)	2.21 (0.03)	0.73 (0.47)
IS(1)	62.08 (0.00)	36.54 (0.00)	25.39 (0.00)	5.98 (0.31)
HR(1)	1.31 (0.19)	4.77 (0.00)	1.72 (0.09)	1.38 (0.17)
$Q(2)$	73.51 (0.00)	42.42 (0.00)	6.31 (0.04)	7.37 (0.03)
LM(2)	3.89 (0.00)	−6.47 (0.00)	−1.33 (0.18)	−2.21 (0.03)
IS(2)	72.63 (0.00)	56.46 (0.00)	27.74 (0.01)	13.29 (0.15)
IS(all)	77.89 (0.00)	69.63 (0.00)	36.31 (0.13)	16.02 (0.38)

The numbers in parentheses represent p-values.
(1): `xtreg n w k ys yr*, vce(cluster id) fe`
(2): `xtreg n w k ys yr* id#c.year, vce(cluster id) fe`
(3): `xtreg D.n D.(w k ys) i.year, vce(cluster id) fe`
(4): `xtreg D.n DL(0/2).(w k ys) i.year, vce(cluster id) fe`

In column (2), I refit the model with firm-specific trends. The coefficient estimates have changed considerably. The wage elasticity has increased (in absolute value) by 25%, and the elasticity with respect to industry output is almost twice as large as before. The residuals still appear to be serially correlated. The four statistics above the dashed line all point toward the presence of first-order correlation. The $Q(2)$ and IS(2) tests indicate there is serial correlation up to the second order as well; the LM(2) test suggests there is correlation of the second order (it does not say anything about first-order correlation, unlike the previous two).

7. An example would be stock data, which can be stable for years and suddenly enter a turbulent phase. Figure 3 in the appendix provides an example of a time series whose variance changes over time.
8. I limit this to second-order serial correlation for the sake of brevity.

Finally, the unrestricted IS test [IS(all)] tests for serial correlation of any order, without limiting the lag length at which it might occur. In this sense, it is less arbitrary than the previous tests, which require the researcher to indicate a cutoff point. Unfortunately, it can be used only in panels that are much wider than they are long and is less powerful than the restricted tests (see section 6).[9]

An alternative option is to fit the model in differences [column (3)]. Again, the coefficients have changed. The elasticity of earnings on employment is now almost −0.5, and the coefficient on industrial output has increased by another 37%. The serial correlation statistics look a lot better. The unrestricted IS statistic at the bottom of the table no longer rejects the null of no serial correlation (at any order). The other tests show more mixed signals, with p-values skirting on both edges of the common 0.05 rejection threshold. Still, we cannot say with confidence that our estimation is free of serial correlation.

Finally, in column (4) we add two lags of the independent variables.[10] I do not introduce lags of the dependent variable for simplicity, because such (short) dynamic panels cannot be estimated consistently with standard fixed-effects regressions.[11] The values presented in the top half of the table refer to the sum of the coefficient estimates and are in line with the previous results, although only the coefficient on capital remains significant.

The residuals seem fairly free of autocorrelation. The tests above the dashed line all point to a lack of first-order serial correlation. The presence of correlation at the second order is less clear, with the $Q(2)$ and LM(2) tests still rejecting the null at 5%, unlike the IS(2) test. The unrestricted IS test points to a complete absence of serial correlation.

Is this last set of estimates more credible than those presented in columns (1)–(3)? Serial correlation statistics by themselves cannot answer that question. However, they are definitely one indicator that these coefficients are less vulnerable to model misspecification than the others.

5 Usage guidelines

The various tests are frequently in disagreement with one another, as can be seen in the example in section 4. Below I provide a few guidelines that might help to reach an informed conclusion in such situations. In section 6, I provide Monte Carlo evidence to support the rules of thumb presented here. Note that all three tests assume that the

9. As a rule of thumb, the unrestricted IS test requires that $N > T^2/2$, where N is the number of panel units and T is the number of time periods. The equivalent requirement for the restricted IS test is $N > p \times T$, where p is the order up to which to test. Section 2 provides more details.
10. The choice of the number of lags is based on the serial correlation tests. In other cases, theory is likely to be the prime guideline. Here we have only a short panel of yearly data, so there are few cues from theory to guide the lag selection (does capital lead to changes in employment with one year's delay? two years'? five years'?).
11. The biggest contribution of the earlier mentioned Arellano and Bond (1991) article is that it introduces a generalized method of moments estimator that can do so.

idiosyncratic errors are uncorrelated with the regressors.[12] This means that including predetermined or endogenous regressors might lead to unreliable test statistics, although the Monte Carlo evidence suggests this is not always the case.

- Use `xtistest`
  - if the data contain gaps (the IS test accommodates them without requiring further action).
  - if you are concerned about higher-order serial correlation, common when individual time trends should be included but are not.
  - only if the panel dimension (N) is significantly larger than the time dimension (T).[13]
- Use `xthrtest`
  - if you suspect the variance changes significantly over time, for example, when studying stock indexes that can be stable for years and turn volatile when a shock happens.
  - only if you are just concerned about first-order serial correlation.
- `xtqptest` is the most reliable in all other situations.
  - The power of the $Q(p)$ (the `lags()` option) increases with N and T, making it suitable for situations where either N or T or both are small. It also performs best in the Monte Carlo simulations (see section 6).
  - The $Q(p)$ test can be complemented with the LM(k) test (the `order` option) to identify at which order the serial correlation is present. For example, if the $Q(2)$ test for serial correlation up to the second order is near the edge of rejection or acceptance territory but the other tests cannot reject the null, then it can be interesting to use the LM(1) and LM(2) tests to check for, respectively, first- and second-order autocorrelation separately.

6 Monte Carlo

In this section, we examine the finite sample performance of the four tests presented in this article and compare them with the existing WD (`xtserial`) and Arellano–Bond (`abar`) commands. The tests are evaluated under seven different scenarios, inspired by issues the practitioner might face. In each scenario, we investigate whether the tests detect serial correlation if it is present (their "power") and accept the null of no serial correlation if there is indeed none (their "size").

12. For `xtqptest` and `xthrtest`, the actual assumption made is a bit more involved, but for all practical purposes, it comes down to the same thing.
13. As a rule of thumb, N should be larger than pT, where N is the number of panel units, T the time length, and p the order up to which you want to test for serial correlation (the `lags()` option).

6.1 Setup

The basic framework defining the panel is described in (10)–(13). Size results are based on $u_{it} = f(\epsilon_{it})$, where ϵ_{it} follows a standard normal distribution. Power results are based on an AR(2) process, such that $u_{it} = f(v_{it}, \epsilon_{it})$, where $v_{it} = 0.1 \times v_{i,t-1} + 0.05 \times v_{i,t-2} + \epsilon_{it}$.[14] Both size and power might depend on the dimension of the panel. To that extent, we set the number of panel units N to 20, 50, or 500 and the time length T to 7 or 50. This way, we capture both typical macro panels (for example, $N = 20$, $T = 50$) and representative micro panels ($N = 500$, $T = 7$) as well as some intermediate forms.

$$y_{it} = c_i + x_{it}\beta + u_{it} \tag{10}$$
$$c_i \sim N(5, 10) \tag{11}$$
$$x_{it} \sim N(0, 1) \tag{12}$$
$$\beta = 1 \tag{13}$$

Table 2 provides a summary of the various scenarios. In the baseline scenario [1], the panel is balanced, and the variance is constant over time. We apply the fixed-effect estimator to obtain the residuals. In scenario [2], the error terms are multiplied by $h_t = e^{0.2t}$, causing the variance to explode over time.[15] Time-varying volatility is particularly common in the financial literature. In scenario [3], the variance is constant once again, but now the panel is unbalanced, with individual panels missing up to a fifth of their first and last observations (uniformly distributed).

Table 2. Scenarios used in Monte Carlo simulations

	Extension
[1] Constant variance	Baseline
[2] Temporal heteroskedasticity	$u_{it} = \epsilon_{it} \times h_t$ and $h_t = e^{0.2t}$
[3] Unbalanced panel	$T_i \neq T_j$ (but no gaps)
[4] No fixed effects	$c_i = 0 = c_j$
[5] Heterogeneous slopes	$\beta_i \sim U[0, 2]$
[6] Dynamic panel	$y_{it} = 0.5y_{i,t-1} + 0.25y_{i,t-2} + c_i + x_{it}\beta + u_{it}$
[7] Cross-sectional dependence	$u_{it} = \lambda_i \times v_t + \epsilon_{it}$ and $\lambda_i \sim U[-1, 3]$ and $v_t \sim N(0, 1)$

In scenario [4], we set all fixed effects to zero. Such a situation might occur when working with first-differenced data or in a panel regression without error component. Scenario [5] imposes panel-specific β_i coefficients (uniformly distributed over $[0, 2]$). We fit this heterogeneous slope model using the mean-group estimator (Pesaran and Smith 1995).

14. In the appendix, I show power results for an MA(1) process.
15. Figures 2 and 3 in the appendix show example time series with constant and time-varying variances.

The final two scenarios put the serial correlation statistics to the test when their assumptions are not met. In scenario [6], y_{it} itself follows an AR(2) process with $y_{it} = 0.5y_{i,t-1} + 0.25y_{i,t-2} + c_i + x_{it}\beta + u_{it}$. Fixed-effects estimation no longer consistently estimates each coefficient (Nickell 1981), one of the requirements of the serial correlation tests.

Scenario [7] introduces cross-sectional dependence into the error structure by introducing a common factor with panel-specific factor loadings [see (14)–(16)].

$$u_{it} = \lambda_i \times v_t + \epsilon_{it} \tag{14}$$

$$\lambda_i \sim U[-1, 3] \tag{15}$$

$$v_t \sim N(0, 1) \tag{16}$$

This cross-sectional correlation violates the assumption of independently distributed errors. Comovement across panel units is a growing concern, especially in the macroeconomic literature (for example, Eberhardt, Helmers, and Strauss [2013]; Chudik, Pesaran, and Tosetti [2011]; Allegretto, Dube, and Reich [2010]). Note that in this specification, fixed-effects estimation is still consistent because the error term remains uncorrelated with the regressors (Eberhardt, Banerjee, and Reade 2010).

6.2 Results

Results for scenarios [1]–[3] are presented in tables 3–4. The LM(1), HR, and WD tests look for first-order correlation only. LM(2) tests look just for second-order correlation. Three tests—IS(2), $Q(2)$, and second-order Arellano–Bond [AB(2)]—look for serial correlation up to the second order. Finally, I also test for serial correlation up to the fourth order with $Q(4)$ to investigate whether the power of the $Q(p)$ test declines if you allow for longer serial correlation than is actually present in the data.

Table 3. Monte Carlo results, scenario [1]–[3], size

	$T=7$			$T=50$		
	$N=20$	$N=50$	$N=500$	$N=20$	$N=50$	$N=500$
No serial correlation						Ideal: 0.05
[1] Constant variance						
- IS(2)	0.00	0.04	0.05	.	.	0.04
- $Q(2)$	0.11	0.06	0.05	0.09	0.07	0.05
- $Q(4)$	0.20	0.09	0.05	0.15	0.09	0.05
- LM(1)	0.08	0.05	0.05	0.06	0.06	0.05
- LM(2)	0.07	0.06	0.04	0.07	0.07	0.06
- HR	0.08	0.06	0.05	0.06	0.06	0.05
- AB(2)	0.43	0.79	1.00	0.09	0.19	0.88
- WD	0.07	0.06	0.06	0.05	0.06	0.05
[2] Time-varying variance						
- IS(2)	0.00	0.39	1.00	.	.	1.00
- $Q(2)$	0.09	0.06	0.08	0.10	0.06	0.07
- $Q(4)$	0.18	0.08	0.10	0.17	0.10	0.14
- LM(1)	0.07	0.06	0.09	0.08	0.06	0.05
- LM(2)	0.07	0.07	0.23	0.06	0.05	0.05
- HR	0.06	0.06	0.05	0.07	0.05	0.05
- AB(2)	0.42	0.80	1.00	0.39	0.40	0.70
- WD	0.09	0.14	0.82	0.13	0.27	0.99
[3] Unbalanced panel						
- IS(2)	0.00	0.05	0.06	.	.	0.04
- $Q(2)$	0.10	0.06	0.04	0.09	0.07	0.06
- $Q(4)$	0.18	0.09	0.04	0.15	0.10	0.06
- LM(1)	0.07	0.06	0.05	0.06	0.06	0.05
- LM(2)	0.07	0.05	0.04	0.08	0.06	0.05
- HR	0.07	0.05	0.05	0.07	0.06	0.05
- AB(2)	0.50	0.86	1.00	0.10	0.21	0.93
- WD	0.07	0.06	0.05	0.06	0.06	0.05

Simulations performed with 2,000 replications.

Table 4. Monte Carlo results, scenario [1]–[3], power

	$T = 7$			$T = 50$		
	$N = 20$	$N = 50$	$N = 500$	$N = 20$	$N = 50$	$N = 500$
AR(2) ($\alpha_1 = 0.10, \alpha_2 = 0.05$)						Ideal: 1.00
[1] Constant variance						
- IS(2)	0.00	0.08	0.77	.	.	1.00
- $Q(2)$	0.14	0.18	0.96	0.85	1.00	1.00
- $Q(4)$	0.21	0.18	0.91	0.83	1.00	1.00
- LM(1)	0.12	0.20	0.94	0.85	1.00	1.00
- LM(2)	0.07	0.07	0.07	0.40	0.77	1.00
- HR	0.09	0.06	0.23	0.73	0.99	1.00
- AB(2)	0.19	0.36	1.00	0.69	0.97	1.00
- WD	0.09	0.10	0.37	0.19	0.35	1.00
[2] Time-varying variance						
- IS(2)	0.00	0.54	1.00	.	.	1.00
- $Q(2)$	0.13	0.20	0.98	0.34	0.66	1.00
- $Q(4)$	0.20	0.20	0.96	0.37	0.58	1.00
- LM(1)	0.11	0.14	0.82	0.32	0.61	1.00
- LM(2)	0.07	0.07	0.11	0.14	0.25	0.98
- HR	0.08	0.07	0.22	0.18	0.33	1.00
- AB(2)	0.21	0.38	1.00	0.64	0.82	1.00
- WD	0.07	0.07	0.22	0.06	0.09	0.43
[3] Unbalanced panel						
- IS(2)	0.00	0.09	0.61	.	.	1.00
- $Q(2)$	0.12	0.14	0.85	0.76	0.99	1.00
- $Q(4)$	0.20	0.15	0.76	0.73	0.98	1.00
- LM(1)	0.10	0.15	0.83	0.75	0.99	1.00
- LM(2)	0.07	0.06	0.05	0.33	0.65	1.00
- HR	0.09	0.06	0.13	0.59	0.93	1.00
- AB(2)	0.27	0.56	1.00	0.53	0.88	1.00
- WD	0.08	0.09	0.31	0.16	0.29	0.99

Simulations performed with 2,000 replications.

Do the tests falsely reject the null (no serial correlation) when there actually is no serial correlation (table 3)?[16] In the baseline scenario [1], with a constant variance and a balanced panel, we find that all tests are decently sized, apart from the Arellano–Bond test, which is not suited for fixed-effects models. Note that the IS(2) test is not valid for two of the dimension combinations. When we introduce temporal heteroskedasticity in scenario [2], both the IS test and the WD test become oversized, worsening as the sample gets larger. The HR test honors its name and stays very close to the optimal size of 0.05 regardless of N and T. The other tests perform worse than in the constant variance case, but the distortions remain relatively minor overall. Moving to an unbalanced panel in scenario [3], we find no significant differences with the balanced case.

As for the power results (do they reject the null when it is indeed false?) in table 4, we find that in the baseline scenario [1], the Q tests have noticeably more power than the standard WD test. The LM(2) and HR tests have fairly little power when $T = 7$, though they improve considerably as the panel gets longer.[17] Introducing time-variation in the variance [2] leads to relatively similar results. The WD test loses even more power, and the Q tests still perform best overall. Surprisingly, the HR test, which was developed for this scenario, has fairly little power, although it does improve as the sample gets bigger in both dimensions. As with the size results, the balance of the sample [3] does not have a large impact on the power of the tests. The minor reductions in power across the board can be attributed to the reduced effective time length of the individual panels (between $0.6 \times T$ and T).

Results for the more challenging scenarios [4]–[7] are presented in tables 5–6. Given the bad performance of the Arellano–Bond test, we omit it in these tables to reduce the number clutter.

16. As is common in the literature, we reject at $p = 0.05$. Thus, a “correctly sized test” will reject the null hypothesis 5% of the time when it is true.
17. It is unclear what to make of the Arellano–Bond power results, given that it was so oversized in these scenarios.

Table 5. Monte Carlo results, scenario [4]–[7], size

	$T = 7$			$T = 50$		
	$N = 20$	$N = 50$	$N = 500$	$N = 20$	$N = 50$	$N = 500$
No serial correlation						Ideal: 0.05
[4] No fixed effects						
- IS(2)	0.00	0.49	1.00	.	.	0.28
- $Q(2)$	0.10	0.07	0.06	0.10	0.07	0.05
- $Q(4)$	0.18	0.10	0.06	0.16	0.09	0.05
- LM(1)	0.07	0.05	0.06	0.07	0.06	0.05
- LM(2)	0.06	0.06	0.05	0.07	0.06	0.05
- HR	0.07	0.06	0.05	0.08	0.05	0.06
- WD	0.07	0.05	0.04	0.05	0.05	0.05
[5] Heterogeneous slopes						
- IS(2)	0.00	0.04	0.06	.	.	0.03
- $Q(2)$	0.09	0.07	0.06	0.09	0.07	0.05
- $Q(4)$	0.17	0.10	0.05	0.16	0.09	0.06
- LM(1)	0.07	0.05	0.05	0.07	0.06	0.05
- LM(2)	0.06	0.05	0.05	0.07	0.06	0.05
- HR	0.06	0.05	0.05	0.07	0.05	0.06
- WD	0.07	0.05	0.05	0.06	0.05	0.05
[6] Dynamic panel						
- IS(2)	0.02	0.07	0.68	.	.	0.07
- $Q(2)$	0.08	0.09	0.90	0.05	0.07	0.61
- $Q(4)$	0.08	0.05	0.79	0.13	0.10	0.69
- LM(1)	0.05	0.09	0.90	0.03	0.03	0.36
- LM(2)	0.06	0.05	0.09	0.06	0.08	0.49
- HR	0.05	0.04	0.04	0.03	0.02	0.06
- WD	0.97	1.00	1.00	1.00	1.00	1.00
[7] Cross-sectional dependence						
- IS(2)	0.00	0.90	1.00	.	.	1.00
- $Q(2)$	0.72	0.88	0.98	0.79	0.93	0.99
- $Q(4)$	0.83	0.96	1.00	0.90	0.99	1.00
- LM(1)	0.58	0.72	0.91	0.64	0.76	0.92
- LM(2)	0.57	0.71	0.91	0.62	0.77	0.93
- HR	0.55	0.72	0.91	0.63	0.77	0.92
- WD	0.58	0.73	0.91	0.64	0.77	0.94

Simulations performed with 2,000 replications.

Table 6. Monte Carlo results, scenario [4]–[7], power

	$T = 7$			$T = 50$		
	$N = 20$	$N = 50$	$N = 500$	$N = 20$	$N = 50$	$N = 500$
AR(2) $(\alpha_1 = 0.10, \alpha_2 = 0.05)$						Ideal: 1.00
[4] No fixed effects						
- IS(2)	0.00	0.80	1.00	.	.	1.00
- $Q(2)$	0.13	0.19	0.96	0.87	1.00	1.00
- $Q(4)$	0.22	0.19	0.91	0.82	0.99	1.00
- LM(1)	0.13	0.20	0.94	0.86	1.00	1.00
- LM(2)	0.06	0.06	0.07	0.41	0.76	1.00
- HR	0.07	0.07	0.25	0.73	0.97	1.00
- WD	0.08	0.10	0.39	0.20	0.36	1.00
[5] Heterogeneous slopes						
- IS(2)	0.00	0.07	0.57	.	.	1.00
- $Q(2)$	0.11	0.15	0.86	0.86	1.00	1.00
- $Q(4)$	0.20	0.15	0.76	0.81	0.99	1.00
- LM(1)	0.11	0.16	0.83	0.85	0.99	1.00
- LM(2)	0.07	0.06	0.07	0.40	0.74	1.00
- HR	0.07	0.06	0.18	0.72	0.97	1.00
- WD	0.08	0.10	0.39	0.20	0.36	1.00
[6] Dynamic panel						
- IS(2)	0.02	0.11	0.98	.	.	1.00
- $Q(2)$	0.10	0.19	1.00	0.69	0.99	1.00
- $Q(4)$	0.09	0.10	0.99	0.63	0.96	1.00
- LM(1)	0.08	0.18	1.00	0.53	0.94	1.00
- LM(2)	0.06	0.04	0.08	0.51	0.89	1.00
- HR	0.05	0.04	0.04	0.35	0.74	1.00
- WD	0.99	1.00	1.00	1.00	1.00	1.00
[7] Cross-sectional dependence						
- IS(2)	0.00	0.93	1.00	.	.	1.00
- $Q(2)$	0.73	0.91	0.99	0.79	0.92	0.99
- $Q(4)$	0.84	0.98	1.00	0.92	0.99	1.00
- LM(1)	0.59	0.72	0.92	0.59	0.75	0.93
- LM(2)	0.58	0.74	0.91	0.62	0.78	0.93
- HR	0.58	0.73	0.92	0.60	0.78	0.92
- WD	0.62	0.76	0.92	0.66	0.78	0.93

Simulations performed with 2,000 replications.

Regarding table 5, in scenario [4], we set the fixed effects to zero. The IS test becomes highly oversized, rejecting the null of no serial correlation far more often than it should (28–100% of the time rather than the ideal 5%). The other statistics are unaffected. Panel-specific coefficients [5] do not appear to trouble any of the tests. On the other hand, including lags of the dependent variable [6] is detrimental to the size of all tests considered, apart from the HR statistic. Performance remains acceptable in smaller panels but gets progressively worse as N and T increase. The WD test completely breaks down.[18] Introducing cross-sectional dependence [7] leads to even worse size results. Every test is now majorly oversized. To my knowledge, there is no serial correlation test that is robust to cross-sectional dependence.[19]

Finally, we move to the power results in table 6. As with the sizes in the previous table, setting the fixed effects to zero [4] does not affect the power of any of the tests, except for the IS test. The introduction of heterogeneous slopes [5] has no discernible impact on the power of the test statistics. Turning to the dynamic panel scenario [6], we find that only the small-sample results are meaningful, given that all the tests became oversized as the sample grew in either dimension. We also find that the power of the tests holds in these small samples. The values presented for the cross-sectional dependence case [7] are entirely meaningless, given that none of the tests "accepted" the null hypothesis of no serial correlation even when it was true.

All in all, we find that, at least in the common scenarios we considered, the Q test is the most robust overall. It has decent size and is generally more powerful than the alternatives, especially compared with the standard WD test.

7 Conclusion

I believe that serial correlation tests can help researchers understand their data and correctly specify their empirical models. Currently, however, serial correlation testing is far from accessible with panel data. Thus, serial correlation testing is often neglected altogether or postponed until the very end of the research cycle. In contrast, the three tests I introduced are easy to use and interpret, making it feasible to test for serial correlation in earlier, exploratory phases of research. Moreover, I provided Monte Carlo evidence indicating they have considerably more power in finite samples than the current generation of tests.

18. In a certain sense, these bad results are inevitable. The bias induced by the lagged dependent variables mainly affects the autoregressive (AR) coefficients. These incorrectly estimated AR parameters in turn lead to serial correlation in the residuals (even if the true errors are serially uncorrelated), which is then detected by the test statistics. In table 9 in the appendix, we introduce endogeneity by correlating the regressor x_{it} with the error term as an alternative specification. Then, the tests are oversized for small samples but approach 5% as the sample gets longer.
19. Table 9 in the appendix shows results for the cross-sectional dependence scenario when we fit the model with the common correlated effects pooled (CCEP) estimator (Pesaran 2006) instead of standard fixed effects. The CCEP estimator filters out the common trends by including the cross-sectional mean of all variables as regressors with panel unit-specific coefficients. We find decent-size results for all tests in that case.

8 References

Allegretto, S., A. Dube, and M. Reich. 2010. Do minimum wages really reduce teen employment? Accounting for heterogeneity and selectivity in state panel data. Institute for Research on Labor and Employment, Working Paper Series from Institute of Industrial Relations, UC Berkeley. https://econpapers.repec.org/paper/cdlindrel/qt7jq2q3j8.htm.

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

Born, B., and J. Breitung. 2016. Testing for serial correlation in fixed-effects panel data models. *Econometric Reviews* 35: 1290–1316.

Choodari-Oskooei, B., and T. P. Morris. 2016. Quantifying the uptake of user-written commands over time. *Stata Journal* 16: 88–95.

Chudik, A., M. H. Pesaran, and E. Tosetti. 2011. Weak and strong cross-section dependence and estimation of large panels. *Econometrics Journal* 14: C45–C90.

Drukker, D. M. 2003. Testing for serial correlation in linear panel-data models. *Stata Journal* 3: 168–177.

Eberhardt, M., A. Banerjee, and J. J. Reade. 2010. Panel estimation for worriers. Economics Series Working Papers 514, University of Oxford, Department of Economics. https://econpapers.repec.org/paper/oxfwpaper/514.htm.

Eberhardt, M., C. Helmers, and H. Strauss. 2013. Do spillovers matter when estimating private returns to R&D? *Review of Economics and Statistics* 95: 436–448.

Inoue, A., and G. Solon. 2006. A Portmanteau test for serially correlated errors in fixed effects models. *Econometric Theory* 22: 835–851.

Nickell, S. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49: 1417–1426.

Pesaran, M. H. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica* 74: 967–1012.

Pesaran, M. H., and R. P. Smith. 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68: 79–113.

Pullenayegum, E. M., R. W. Platt, M. Barwick, B. M. Feldman, M. Offringa, and L. Thabane. 2016. Knowledge translation in biostatistics: A survey of current practices, preferences, and barriers to the dissemination and uptake of new statistical methods. *Statistics in Medicine* 35: 805–818.

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

Solon, G. 1984. The effects of unemployment insurance eligibility rules on job quitting behavior. *Journal of Human Resources* 19: 118–126.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

About the author

Jesse Wursten is a PhD student at KU Leuven.

A Appendix

A.1 Example time series

Figures 2–3 illustrate the difference between residuals with a constant variance and those whose variance increases exponentially over time. The top panel shows an example time series without serial correlation, the bottom panel one with mild first- and second-order autocorrelation.

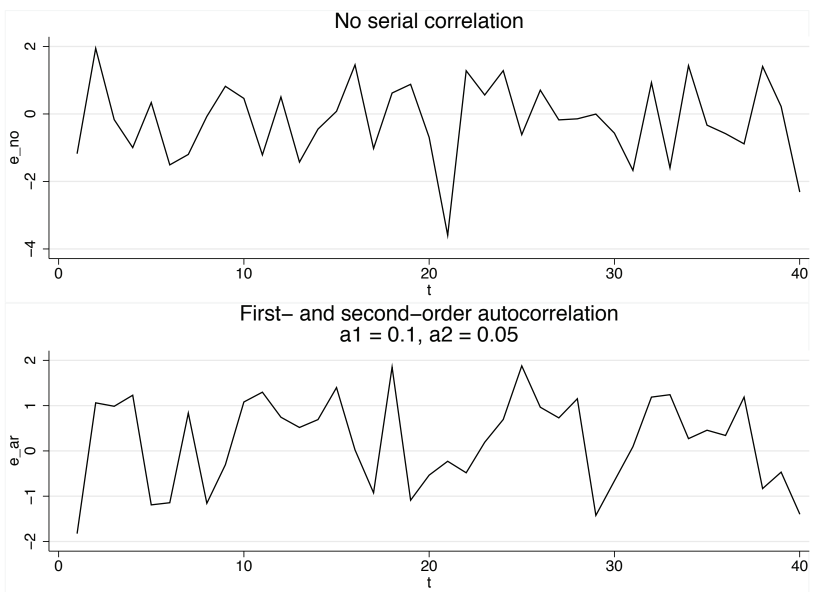

Figure 2. Constant variance, with and without serial correlation

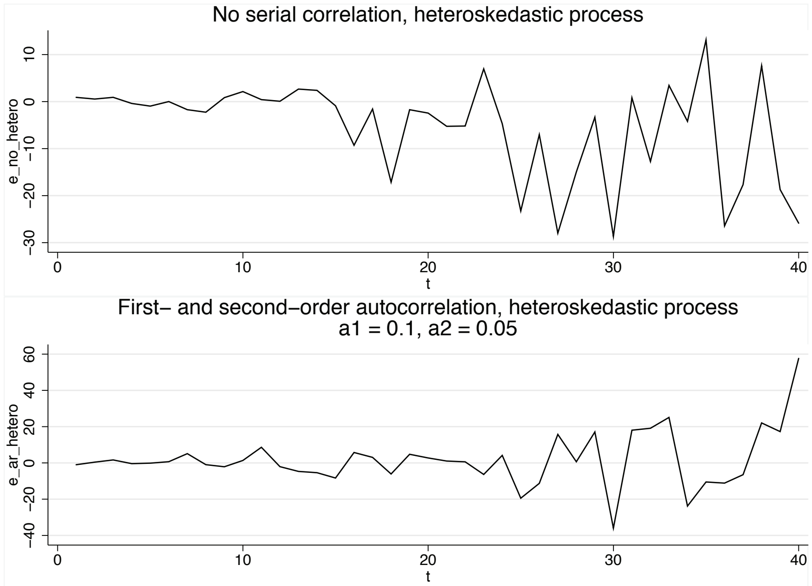

Figure 3. Time-varying ("heteroskedastic") variance, with and without serial correlation

A.2 Monte Carlo evidence: Moving averages

Tables 7–8 present power results for errors following a first-order moving-average [MA(1)] process, with $\theta_1 = 0.10$. There are no major differences compared with the AR(2) results, with the Q test performing best overall. We include the unrestricted IS test in these tables. Overall, its performance is very similar to that of the IS(2) test, with a mild reduction in power, but note that in many cases, it cannot be estimated because the dimension of the test is too large.

Table 7. Monte Carlo results, MA(1)-process, scenario [1]–[3], power

	$T=5$			$T=50$		
	$N=20$	$N=50$	$N=500$	$N=20$	$N=50$	$N=500$
MA(1) ($\theta_1 = 0.10$)						Ideal: 1.00
[1] Constant variance						
- IS(2)	0.00	0.08	0.89	.	.	1.00
- IS(all)	0.00	0.07	0.80	.	.	.
- $Q(2)$	0.16	0.25	0.99	0.81	1.00	1.00
- $Q(4)$	0.23	0.23	0.97	0.78	0.99	1.00
- LM(1)	0.14	0.30	0.99	0.86	1.00	1.00
- LM(2)	0.08	0.11	0.53	0.07	0.07	0.11
- HR	0.10	0.17	0.89	0.81	0.99	1.00
- AB(2)	0.13	0.29	1.00	0.66	0.97	1.00
- WD	0.14	0.21	0.95	0.59	0.95	1.00
[2] Time-varying variance						
- IS(2)	0.00	0.52	1.00	.	.	1.00
- IS(all)	0.00	0.35	1.00	.	.	.
- $Q(2)$	0.15	0.26	0.99	0.30	0.54	1.00
- $Q(4)$	0.24	0.23	0.98	0.34	0.48	1.00
- LM(1)	0.12	0.23	0.97	0.32	0.61	1.00
- LM(2)	0.11	0.17	0.90	0.07	0.06	0.06
- HR	0.10	0.16	0.86	0.26	0.51	1.00
- AB(2)	0.15	0.31	1.00	0.61	0.81	1.00
- WD	0.07	0.06	0.11	0.07	0.07	0.16
[3] Unbalanced panel						
- IS(2)	0.01	0.08	0.82	.	.	1.00
- IS(all)	0.01	0.08	0.72	.	.	.
- $Q(2)$	0.13	0.20	0.96	0.72	0.98	1.00
- $Q(4)$	0.21	0.21	0.92	0.68	0.96	1.00
- LM(1)	0.12	0.24	0.98	0.79	0.99	1.00
- LM(2)	0.09	0.11	0.59	0.07	0.07	0.11
- HR	0.09	0.12	0.73	0.70	0.98	1.00
- AB(2)	0.20	0.46	1.00	0.50	0.91	1.00
- WD	0.12	0.18	0.89	0.50	0.90	1.00

Simulations performed with 2,000 replications.

Table 8. Monte Carlo results, MA(1)-process, scenario [4]–[7], power

	$T=7$			$T=50$		
	$N=20$	$N=50$	$N=500$	$N=20$	$N=50$	$N=500$
MA(1) ($\theta_1 = 0.10$)						Ideal: 1.00
[4] No fixed effects						
- IS(2)	0.00	0.73	1.00	.	.	1.00
- IS(all)	0.00	0.63	1.00	.	.	.
- $Q(2)$	0.15	0.24	0.99	0.80	0.99	1.00
- $Q(4)$	0.23	0.22	0.97	0.77	0.98	1.00
- LM(1)	0.15	0.30	0.99	0.86	1.00	1.00
- LM(2)	0.09	0.11	0.53	0.07	0.06	0.10
- HR	0.10	0.17	0.88	0.80	0.99	1.00
- WD	0.13	0.22	0.95	0.59	0.95	1.00
[5] Heterogeneous slopes						
- IS(2)	0.00	0.06	0.71	.	.	1.00
- IS(all)	0.00	0.05	0.58	.	.	.
- $Q(2)$	0.12	0.18	0.94	0.79	0.99	1.00
- $Q(4)$	0.21	0.18	0.88	0.75	0.98	1.00
- LM(1)	0.11	0.22	0.96	0.85	1.00	1.00
- LM(2)	0.07	0.09	0.40	0.07	0.07	0.10
- HR	0.08	0.13	0.73	0.80	0.99	1.00
- WD	0.13	0.22	0.95	0.59	0.95	1.00
[6] Dynamic panel						
- IS(2)	0.02	0.10	0.98	.	.	1.00
- IS(all)	0.02	0.10	0.92	.	.	.
- $Q(2)$	0.10	0.20	1.00	0.37	0.83	1.00
- $Q(4)$	0.09	0.12	1.00	0.39	0.74	1.00
- LM(1)	0.09	0.26	1.00	0.48	0.91	1.00
- LM(2)	0.07	0.11	0.55	0.08	0.11	0.73
- HR	0.05	0.07	0.28	0.35	0.76	1.00
- WD	1.00	1.00	1.00	1.00	1.00	1.00
[7] Cross-sectional dependence						
- IS(2)	0.00	0.93	1.00	.	.	1.00
- IS(all)	0.00	0.81	1.00	.	.	.
- $Q(2)$	0.73	0.89	0.99	0.79	0.93	0.99
- $Q(4)$	0.84	0.97	1.00	0.92	0.99	1.00
- LM(1)	0.56	0.74	0.91	0.62	0.77	0.93
- LM(2)	0.59	0.73	0.91	0.60	0.77	0.93
- HR	0.55	0.72	0.92	0.63	0.76	0.93
- WD	0.62	0.76	0.92	0.67	0.79	0.94

Simulations performed with 2,000 replications.

A.3 Monte Carlo evidence: Endogenous regressor and cross-sectional dependence

If we introduce a bias in the coefficient of the regressor x_{it} by correlating it to the error term, we find that in longer panels, the size still remains acceptable (top half of table 9). This is in contrast with the lagged dependent variable approach in table 5. The good performance of the WD test is due to a lack of serial correlation in the endogenous regressor and may not extend to more realistic situations.

Table 9. Monte Carlo results, endogenous regressor and CDP, size

	$T = 7$			$T = 50$		
	$N = 20$	$N = 50$	$N = 500$	$N = 20$	$N = 50$	$N = 500$
No serial correlation						Ideal: 0.05
Endogenous regressor						
- IS(2)	0.00	1.00	1.00	.	.	0.76
- $Q(2)$	0.43	0.61	0.94	0.13	0.09	0.13
- $Q(4)$	0.73	0.85	1.00	0.21	0.12	0.16
- LM(1)	0.28	0.41	0.76	0.09	0.08	0.11
- LM(2)	0.28	0.39	0.76	0.09	0.07	0.10
- HR	0.27	0.39	0.76	0.10	0.07	0.11
- WD	0.07	0.06	0.05	0.05	0.05	0.05
Cross-sectional dependence						
Estimated with CCEP-estimator						
- IS(2)	0.00	0.04	0.05	.	.	0.03
- $Q(2)$	0.10	0.07	0.05	0.09	0.06	0.05
- $Q(4)$	0.22	0.10	0.06	0.17	0.09	0.05
- LM(1)	0.07	0.05	0.05	0.07	0.06	0.05
- LM(2)	0.07	0.05	0.05	0.07	0.05	0.05
- HR	0.07	0.06	0.04	0.07	0.06	0.05

Simulations performed with 2,000 replications.

In the bottom half of table 9, we see that if we estimate our cross-sectionally dependent panel with the common correlated effects pooled (CCEP) estimator (Pesaran 2006), sizes of all tests return to normal. The CCEP estimator adds the cross-sectional averages of each variable (dependent and independent) as regressor, with cross section-specific coefficients. This filters out the cross-sectional dependence if N is sufficiently large.