NBER TECHNICAL PAPER SERIES

TESTING FOR INDIVIDUAL EFFECTS IN DYNAMIC MODELS USING PANEL DATA

Douglas Holtz-Eakin

Technical Working Paper No. 57

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 1986

The research reported here is part of the NBER's research program in Taxation. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

Technical Working Paper #57
June 1986

Testing for Individual Effects in Dynamic Models Using Panel Data

ABSTRACT

This note presents a simple, linear test for individual effects in dynamic models using panel data; building upon the techniques of Holtz-Eakin, Newey, and Rosen (HNR) [1985] for estimating vector autoregressions using panel data.

While implementing estimators which are consistent in the presence of individual effects is straightforward, there is no guarantee that this form of heterogeneity is an important feature of the data. Moreover, there are advantages to avoiding an individual effects specification. Thus, it is useful to have a test for the existence of individual effects.

The test focuses on sample moment conditions implied by the presence of individual effects and is particularly suited for dynamic models using panel data. The calculations follow directly from linear, instrumental variable techniques which are computationally straightforward. Moreover, the test statistics follows directly from the estimation of autoregressive models.

Douglas Holtz-Eakin
Department of Economics
Columbia University
New York, NY 10027

This note presents a simple, linear test for individual effects in dynamic models using panel data. It builds upon the techniques of Holtz-Eakin, Newey, and Rosen (HNR) [1985] for estimating vector autoregressions using panel data. HNR present estimation and testing procedures for dynamic models under the maintained assumption that the data under analysis contain individual-specific effects which are unobserved by the econometrician. Below, this set of procedures is modified to test for the presence of individual effects in the sample under analysis.

The use of panel data sets for empirical analysis of the behavior of individual economic units has become increasingly popular. One advantage of using panel data is that an individual effects specification may be used to summarize cross-sectional heterogeneity which is not observed by the econometrician. The primary complication is that these individual effects may be correlated with the observed data for the right side variables. If such correlation is present, consistent estimation will not be possible using a single cross-section of data and there will be no internal evidence of the problem. Through the use of panel data consistent parameter estimates may be obtained.

While implementing individual effects estimators is straightforward, there is no guarantee that this form of heterogeneity is an important feature of the data. Moreover, there are advantages to avoiding an individual effects specification. First, a single cross-section is sufficient data to obtain consistent estimates of the parameters. If a panel of data is still available all the variation may be used to identify parameters. In contrast, one method to control for individual effects is to formulate the model in deviations from the time means for each unit. This has the drawback that the parameters of time-invariant (but cross-sectionally varying) variables may not be estimated.

Further, the use of deviations from means is innappropriate in dynamic models; i.e. models with lagged values of the dependent variable on the right side. In these models, differencing may be used to remove the individual effects. Here again the parameters of time-invariant variables may not be estimated. In addition, differencing typically increases the relative importance of measurement error in the total variation present; making correct inferences more difficult. In sum, it is useful to have a test for the existence of individual effects.

The test below is particularly suited for dynamic models and follows directly from the techniques proposed by HNR for estimating vector autoregressions using panel data. Alternative tests have been proposed by Chamberlain [1983] and Hausman and Taylor [1981]. Chamberlain's test requires non-linear estimation methods, while the Hausman and Taylor specification test is not appropriate in a dynamic setting. The test below uses linear, instrumental variable techniques which are computationally straightforward. Moreover, the test statistics follows directly from the estimation of autoregressive models.

In the next section I present the basic idea. Section 2 develops the test in greater detail. The third section contains an empirical example of the test in the context of estimating an equation of wage dynamics using a sample of male employees' wages. The final section is a summary with conclusions.

1.) <u>Basic Method</u>

The basic notion is easily seen by considering the following simple autoregressive model:

(1.1) $$y_{it} = \beta y_{it-1} + f_i + \eta_{it}$$

(1.1′) $$y_{it} = \beta y_{it-1} + \epsilon_{it} \qquad i=1..N,\ t=1..T$$

where i indexes individual cross-sectional units, t indexes time periods, and $\epsilon_{it} \equiv f_i + \eta_{it}$. In specifications of static models using panel data the primary complication is the potential for correlation between the right-side variables and the individual effects (f_i's); with the result that OLS leads to inconsistent parameter estimates. As is apparent from equation (1.1), assuming an autoregressive structure ensures that this correlation will occur due to the presence of y_{it-1} on the right side. Put differently, the necessary condition to obtain consistent estimates of β is:

(1.2) $$E\{y_{it-1}\epsilon_{it}\} = 0$$

In the presence of the f_i, the orthogonality condition (1.2) will be violated. Notice also that the use of an instrumental variable, say y_{t-2}, does not solve the problem. In this case, the necessary condition is:

(1.2') $$E\{y_{it-2}\epsilon_{it}\} = 0$$

which is also violated in the presence of fixed effects.

To obtain consistent estimates, first-difference (1.1) to eliminate the individual effects:[1]

(1.3) $$y_{it}-y_{it-1} = \beta(y_{it-1}-y_{it-2}) + \epsilon_{it}-\epsilon_{it-1}$$

and note that the individual effects are eliminated; $\varepsilon_{it}-\varepsilon_{it-1}=\eta_{it}-\eta_{it-1}$. Due to the induced serial correlation in the error term and the presence of lagged dependent variables, equation (1.3) must be estimated using instrumental variables. In this case, candidates for valid instrumental variables are lagged values of y dated t-2 or earlier. Using y_{t-2}, the necessary orthogonality condition is:

(1.4) $$E\{y_{t-2}(\varepsilon_{it}-\varepsilon_{it-1})\} = 0$$

which is will be satisfied both in the presence and in the absence of individual effects.

The orthogonality conditions in equations (1.2) and (1.4) may be exploited to test for the presence of fixed effects. Consider a panel data set which contains observations on the variable y for three periods. The simple AR(1) model in equation (1.1) may be estimated for the last two periods. Under the null hypothesis of no individual effects, the following orthogonality conditions hold:

(1.5a) $$E\{y_{i2}\varepsilon_{i3}\} = 0$$

(1.5b) $$E\{y_{i1}\varepsilon_{i3}\} = 0 \qquad i=1..N$$

(1.5c) $$E\{y_{i1}\varepsilon_{i2}\} = 0$$

Thus, there are three orthogonality conditions available to identify the single parameter. The remaining two overidentifying restrictions may be used to test for individual effects.

To see how, reformulate the conditions (1.5) as:

(1.6a) $$E\{y_{i1}(\varepsilon_{i3}-\varepsilon_{i2})\} = 0$$

(1.6b) $$E\{y_{i1}\varepsilon_{i2}\} = 0$$

(1.6c) $$E\{y_{i2}\varepsilon_{i3}\} = 0$$

Notice that equations (1.6) impose the same restrictions as equations (1.5) and, thus, are equivalent under the null hypothesis. On the other hand, condition (1.6a) will also hold in the presence of fixed effects and can be used to identify a first-differenced estimator of β under the alternative hypothesis of individual effects. The null hypothesis imposes only the two additional restrictions (1.6b) and (1.6c) on the data. Intuitively, the test for individual effects is a test of whether the sample moments corresponding to these restrictions are sufficiently close to zero; contingent upon imposing (1.6a) to identify the parameter β.

To implement the test, first estimate the model under the null hypothesis. It is conceptually simplest to think of different time periods as different "equations" since cross-sectional variation may be used to identify the parameters. Thus, under the null, there is an equation for period 2 and one for period 3. Estimates of the parameter, β, may be obtained using the method of moments. Efficient estimation requires that they be estimated jointly with the restriction imposed that the estimated value of β be the same for both periods. Using the estimated parameter, sample residuals may be calculated and a chi-square test statistic for whether the sample moments satisfy the orthogonality conditions (1.6) computed. (See below for the details of the computations.)

To test for individual effects one needs to measure the marginal contribution of the conditions (1.6b) and (1.6c) after first imposing only (1.6a). Since (1.6a) holds in the presence of fixed effects, this amounts to a comparison of estimation results under the null versus the alternative hypothesis. The only complication is that the equation for period 2 is not estimable in first differenced form (because y_{i0} -- the necessary instrumental variable -- is not observed) and the test must correctly account for the fact that some parameters are not identified under the alternative hypothesis.

As the example makes clear, implementation of the test requires specification of the dynamic structure of the model, consideration of the appropriate orthogonality conditions, and computation of the parameter estimates and test statistics. The next section considers these issues in greater detail.

2.) Estimation and Testing

To approach the problem in more detail, consider a general specification for the autoregressive model:

$$(2.1) \qquad y_{it} = \Sigma\beta_k y_{it-k} + f_i + \eta_{it} \qquad k=1..m$$

where the maximum lag length, m, is unknown.[2] The null hypothesis of no individual effects implies the orthogonality conditions:

$$(2.2) \qquad E\{y_{it-j}\epsilon_{it}\} = 0 \qquad j=1..t-1,\ t=(m+1)...T$$

Notice that only the equations for t>m may be estimated in order to accomodate the lag distribution. As a result, there are T-m equations containing the m parameters. The null hypothesis imposes $R=[T(T-1)-m(m-1)]/2$ orthogonality conditions (ignoring constant terms) which may be used both to identify the parameters of (2.1) and leaves (R-m) overidentifying restrictions which may contribute to a test of the null hypothesis.

As in the example above, the conditions (2.2) may be reformulated in a manner which sheds more light on the test. An equivalent set of R restrictions is:

$$(2.3a) \quad E\{y_{it-j}(\epsilon_{it}-\epsilon_{it-1})\} = 0 \qquad j=2..t-1,\ t=(m+2)..T$$

$$(2.3b) \quad E\{y_{im+1-j}\epsilon_{im+1}\} = 0 \qquad j=1..m$$

$$(2.3c) \quad E\{y_{it-1}\epsilon_{it}\} = 0 \qquad t=(m+2)..T$$

The restrictions (2.3a) are the moment conditions needed to identify the first differenced estimator and hold under both the null and alternative hypotheses. The restrictions (2.3b) are those needed to estimate the equation for period (m+1) in levels using the previous m lags as instrumental variables. Similarly, the restrictions in (2.3c) are those needed to admit the first lag of y as an instrumental variable in the estimation of the last (T-m-1) equations in levels.

Before considering the empirical implementation of these restrictions, it is worth noting that the investigator may not wish to implement the full set of orthogonality conditions. For example, in the empirical work below a panel data set covering 14 years is used (T=14). Choosing even a relatively long lag length, say m=5, gives R=81 orthogonality conditions. Included among these are the covariance between y_{i1} and ϵ_{i14}. Covariances at such long lags are likely to be quite small in practice, contribute little to sample measures of covariance, and lower the power of tests.

Instead, consider imposing the orthogonality conditions on only the most recent q covariances, i.e.:

$$(2.4) \quad E\{y_{it-j}\epsilon_{it}\} = 0 \qquad \begin{array}{l} j=1..q,\ \text{if } q<=t-1 \\ j=1..t-1,\ \text{if } q>t-1 \\ t=(m+1)...T \end{array}$$

In this case, fewer restrictions are imposed, the total number being $R=qT-\{[q(q+1)+m(m-1)]/2\}$.

Once again, a reformulation of the moment conditions is desirable. Here there are two cases. First, when $q<=m$, the conditions are a straightforward modification of those in (2.3):

$$(2.5a) \quad E\{y_{it-j}(\varepsilon_{it}-\varepsilon_{it-1})\} = 0 \qquad j=2..q,\ t=(m+2)..T$$

$$(2.5b) \quad E\{y_{im+1-j}\varepsilon_{im+1}\} = 0 \qquad j=1..q$$

$$(2.5c) \quad E\{y_{it-1}\varepsilon_{it}\} = 0 \qquad t=(m+2)..T$$

If, instead, $q>m$ there are fewer than q conditions imposed on the first (q-m-1) equations, viz.:

$$(2.6a) \quad E\{y_{it-j}(\varepsilon_{it}-\varepsilon_{it-1})\} = 0 \qquad j=2..t-1,\ t=(m+2)..q$$

$$(2.6b) \quad E\{y_{it-j}(\varepsilon_{it}-\varepsilon_{it-1})\} = 0 \qquad j=2..q,\ t=(q+1)..T$$

$$(2.6c) \quad E\{y_{im+1-j}\varepsilon_{im+1}\} = 0 \qquad j=1..m$$

$$(2.6d) \quad E\{y_{it-1}\varepsilon_{it}\} = 0 \qquad t=(m+2)..T$$

In each of these cases, the orthogonality conditions restrict the covariance between candidate instrumental variables and the error terms of two types of models: a first differenced model and the model specified in levels. The former restrictions will hold under both the null and the alternative hypotheses and may be used to identify the parameters of the model. The latter set may be used to test for the existence of individual effects.

To implement estimation and testing, stack the cross-sectional observations for any equation, s, and write the model as:

$$(2.7) \quad Y_s = W_s\beta + \varepsilon_s$$

Notice that there will be more than one equation for most time periods due to the use of both differenced and levels equations under the null hypothesis. For example, in the model from Section 1 the equations corresponding to (2.7) are:

(2.7a) $Y_3-Y_2 = (Y_2-Y_1)\beta + (\varepsilon_3-\varepsilon_2)$

(2.7b) $Y_3 = Y_2\beta + \varepsilon_3$

(2.7c) $Y_2 = Y_1\beta + \varepsilon_2$

Next "stack" the equations (2.7) to form a system. With this, the observations for equations (2.7) may be written:

(2.8) $Y = W\beta + \varepsilon$

The parameters of equation (2.8) may be estimated by instrumental variables where the list of appropriate instrumental variables is given by the orthogonality conditions discussed above. Let Z_s be the matrix of instrumental variables for equation s. Choosing a block diagonal matrix, Z, of instrumental variables (the matrix of instrumental variables for each equation on the diagonal), the orthogonality conditions ensure that:

(2.9) $\operatorname{plim}_{N\to\infty} (Z'\varepsilon)/N = 0$

To estimate β, premultiply (2.8) by Z' to obtain:

(2.10) $Z'Y = Z'W\beta + Z'\varepsilon$

One can then form a consistent instrumental variables estimator by applying GLS to this equation; estimating the covariance matrix, Ω, of the (transformed) disturbances as outlined in HNR. The result is that an estimate of β is efficient in the class of instrumental variable estimators which use linear combinations of the instrumental variables. This follows directly from the results of Hansen [1982]. (See also White [1980].)

To test hypotheses, use the chi-square test in HNR. Let Q be:

$$(2.11) \qquad Q = (Y-W\beta)'Z(\Omega^{-1})Z'(Y-W\beta) \ / \ N$$

Q, the weighted sum of squared residuals, has a chi-square distribution with degrees of freedom equal to the number of overidentifying restrictions as N grows. When comparing two nested estimation results, the appropriate test statistic is:

$$(2.12) \qquad L = Q_R - Q$$

where Q_R is the sum of squared residuals from the restricted estimation. L has a chi-squared distribution with degrees of freedom equal to the degrees of freedom of Q_R minus the degrees of freedom of Q. In this applications, Q_R is the sum of squared residuals when imposing the full set of orthogonality conditions implied by the null hypothesis, and Q is the sum of squared residuals from imposing only those restrictions needed for the first differenced versions.

Importantly, the same estimate of Ω must be used in both computations. Because not all the equations are identified under the alternative hypothesis, Ω is estimated under the null hypothesis. When calculating parameters and test statistics under the alternative hypothesis, the inverse of the submatrix corresponding the the differenced equations only is employed.[3]

3.) Empirical Example

In this section, I demonstrate the test in the context of estimating an equation of wage dynamics. The data used are the log of average annual hourly earnings for 898 males over the years 1968 to 1981.[4]

The goal is to estimate the autoregressive repesentation of wages over time:

$$(3.1) \qquad w_{it} = \Sigma\alpha_k w_{i,t-k} + f_i + \eta_{it} \qquad k=1..m, \; t=(m+1)..T$$
$$= \Sigma\alpha_k w_{i,t-k} + \varepsilon_{it}$$

determining the appropriate lag length and to test the hypothesis of no individual effects.

The tests were conducted in the following fashion. Initially, the model was specified as an AR(7). Notice that this is equivalent to specifying equation (3.1) as an AR(6), i.e. m=6, and differencing to eliminate the f_i. As a result, equations for the years 1975 to 1981 are estimable.

Twice lagged values (w_{t-2}) of the dependent variable for each year were used as an instrumental variables, the equations estimated, and an estimate of the asymptotic covariance matrix Ω constructed. The equations were successively re-estimated until the lag length could no longer be reduced. Since the main danger is to innappropriately truncate the lag distribution (Type II error) these tests were conducted at the 10% level. The result is an AR(2) specification for the reduced form wage equation. The lag length tests are shown in Table 1. This sequence is exactly that proposed by HNR to determine the appropriate lag length in the presence of individual effects.

Given this specification, the test for individual effects was conducted. In doing so, restrictions were placed only on the first two covariances (i.e. q=2) rather than on the covariance with all potential lags. As shown in Table 1, the hypothesis of no fixed effects is strongly rejected. Thus, as is typically assumed, wage determination over time contains an important element of unobserved heterogeneity and the steady state of the wage process will differ across individuals.

Finally, the AR(2) reduced form and the presence of individual effects suggests that the correct model may be an AR(1) in first differences. A formal test rejects this restriction. The result is given in the last line of Table 1. The resulting parameter estimates for the AR(2) wage equation are given in Table 2.

The upshot of this specification procedure is: (i) evidence that individual effects are an important component of wage determination. This supports the conventional practice of using first differences in panel studies of individual wages, and (ii) empirical evidence that the appropriate specification of wage dynamics (conditional on individual effects) is an AR(2).

4.) Summary

This note has demonstrated an extension of the methods of Holtz-Eakin, Newey, and Rosen to testing for the presence of individual effects in dynamic models using panel data. The estimated example finds strong evidence of individual effects in the specification of the dynamics of the wage process using a subsample of the PSID.

Table 1

Chi-Square Test Results[a]

	Log Wage L	DF
Lag 7 = 0	0.07	1
Lag 6 = 0	0.07	1
Lag 5 = 0	0.09	1
Lag 4 = 0	0.29	1
Lag 3 = 0	0.49	1
Lag 2 = 0	2.79[d]	1
No Individual Effects	98.08[b]	9
Lag 2 = 0, In First Differences	6.86[b]	1

[a]Equation estimated for years 1975 to 1981. All equations included dummy variables for each year.

[b]Significant at 1% level.

[c]Significant at 5% level.

[d]Significant at 10% level.

Table 2

Parameter Estimates
(t-ratios)

Lag 1	0.323 (6.457)
Lag 2	0.098 (2.620)

Notes

*I am indebted to Joe Altonji, Bruce Lehmann and Whitney Newey for comments on previous drafts and to Joe Altonji for generously providing the data. I particularly thank Whitney Newey for detecting an error in the original version. I thank the Council for Research in the Humanities and Social Sciences at Columbia University for financial support. This research is part of the National Bureau of Economic Research Project on State and Local Government Finance.

[1] As is well know, taking deviations from individual specific time means (the "within" estimator) will also eliminate the individual effect. However, in the presence of lagged values of the dependent variable, this technique fails to yield consistent estimates. Throughout, I restrict the discussion to the first difference approach.

[2] HNR discuss the inference of m in great detail and allow both the lag parameters and the individual effects to be time varying. Here I restrict the discussion to the case of time-invariant parameters. Also, while not shown, is straightforward to include other variables on the right side of equation (2.1).

[3] HNR discuss this in greater detail, including the estimation of Ω in the presence of heteroskedasticity.

[4] Altonji and Paxson [1986] has a complete discussion of the data.

References

Altonji, J. and C. Paxson, "Job Characteristics and Hours of Work", forthcoming in Research in Labor Economics, 1986.

Chamberlain, G. "Panel Data," Chapter 22 in Z. Griliches and M. Intrilligator (ed.s) The Handbook of Econometrics, Volume II, Amsterdam: North-Holland, 1983.

Hansen, L. "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica, 1982, pp. 1029-1054.

Hausman, J. and W. Taylor (1981), "Panel Data and Unobservable Individual Effects", Econometrica, 49, pp. 1377-1398.

Holtz-Eakin, D., W. Newey, and H. Rosen, "Implementing Causality Tests with Panel Data, with an Example from Local Public Finance," National Bureau of Economics Technical Working Paper No. 48, 1985.

White, H. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," Econometrica, 1980, pp. 817-838.