ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 55-56 – 1999

Specification Tests in Panel Data Models Using Artificial Regressions

Badi H. BALTAGI*

ABSTRACT. – This paper surveys some applications of artificial regressions including the GAUSS-NEWTON, Double-Length and Binary Response Model regressions as testing tools for panel data models. In addition, several other artificial regression tests are reviewed including HAUSMAN's [1978] specification test, CHAMBERLAIN's [1982] omnibus goodness-of-fit test and WOOLDRIDGE's [1995] simple variable addition tests for selection bias. The important point to emphasize is that in many cases these artificial regressions provide the easiest way to compute specification tests, and in most cases provide a reasonably easy way to do so.

Tests de spécifications dans les modèles de données de panel à l'aide de régressions artificielles

RÉSUMÉ. – Cet article étudie quelques usages des régressions artificielles telles que le modèle de GAUSS-NEWTON, le modèle « Double-Length » et celui de réponse binaire comme des outils de test pour les modèles de panel. En sus, plusieurs tests basés sur des régressions artificielles sont passés en revue tels que le test de spécification d'HAUSMAN [1978], le test d'adéquation de CHAMBERLAIN [1982] et celui de biais de sélection par addition d'une variable de WOOLDRIDGE [1995]. Le fait que dans de nombreux cas, ces régressions fournissent le moyen le plus simple d'effectuer des tests de spécification et ce souvent de manière raisonnable doit être souligné.

* Badi H. BALTAGI: Department of Economics, Texas A&M University.

I would like to thank P. SEVESTRE and two anonymous reviewers for their helpful comments and suggestions.

1 Introduction

Davidson and MacKinnon [1993] gave numerous applications of specification tests using artificial regressions including the Gauss-Newton regression (GNR), the Double-Length regression (DLR) and the Binary Response Model regression (BRMR). Specific applications in econometrics include tests for serial correlation, heteroskedasticity, non-nested regressions, structural chan-ge, functional form and Hausman type tests based on comparing two sets of estimators. This paper extends the application of these artificial regressions to specification tests using panel data regression models.

In particular, we focus on three specific applications where these tests are useful in computing specification tests for the error component regression model. The first application, given in Section 2, shows that the GNR test for zero random individual effects in the context of a panel data linear regression model reduces to a variable addition test which tests the significance of one additional variable in the original regression. This additional regressor is the vector of least squares residuals summed over time and expressed as deviations from the original residuals. The test-statistic is $1/\sqrt{2}$ times the *t*-statistic on this additional variable. The second application, given in Section 3, shows how the DLR can be used to test for linear and log-linear error component regressions against Box-Cox alternatives. The third application, given in Section 4, tests for fixed effects logit and probit panel data models using a binary response model regression. These applications demonstrate that these artificial regressions provide an easy way to compute specification tests in panel data models. Section 5 of the paper surveys other specification tests in panel data using artificial regressions, while Section 6 provides our conclusion.

2 Testing for Random Individual Effects

Consider the *non-linear* regression model

$$(2.1) \qquad y_{it} = x_{it}(\beta) + u_{it} \qquad i = 1,2,..,N; \; t = 1,2,..,T$$

where β is a $K \times 1$ vector, and $x_{it}(\beta)$ is a scalar (usually non-linear) function of the regressors observed for the *i*-th individual in the *t*-th time period. The disturbances follow a one-way error component model, given by

$$(2.2) \qquad u_{it} = \mu_i + \nu_{it}$$

where $\mu_i \sim \text{IID}\,(0,\sigma_\mu^2)$ denote the random individual (time-invariant) effects and $\nu_{it} \sim \text{IID}\,(0,\sigma_\nu^2)$ denote the remainder effects. These error components are assumed independant of each other among themselves. It is well known that the FULLER and BATTESE [1974] transformation applied to (2.1) gives

$$y^*_{it} = x^*_{it}(\beta) + u^*_{it} \tag{2.3}$$

where $y^*_{it} = (y_{it} - \theta\overline{y}_i)$ with $\overline{y}_{i.} = \sum_{t=1}^{T} y_{it}/T_{it}$. Similarly, $x^*_{it}(\beta) = [x_{it}(\beta) - \theta_{\bar{x}_{i.}}(\beta)]$ with $\overline{x}_{i.}(\beta) = \sum_{t=1}^{T} x_{it}(\beta)/T$. Also, $u^*_{it} = (u_{it} - \theta\bar{u}_{i.})$ with $\overline{u}_{i.} = \sum_{t=1}^{T} u_{it}/T$. In this case, $\theta = 1 - (\sigma_\nu/\sigma_1)$ with $\sigma^2_1 = T\sigma^2_\mu + \sigma^2_\nu$. The transformed disturbances u^*_{it} have mean zero and constant variance σ^2_ν. Least squares on (2.3) is equivalent to generalized least squares on (2.1).

A LAGRANGE-Multiplier test for zero random individual effects in a linear regression model, *i.e.*, H_o; $\sigma^2_\mu = 0$, has been derived by BREUSCH and PAGAN [1980]. This can be extended to the context of the transformed non-linear regression (2.3), as we can rewrite this model as

$$y_{it} = Z_{it}\,(\beta,\theta) + u^*_{it} \tag{2.4}$$

where

$$Z_{it}\,(\beta,\theta) = x^*_{it}(\beta) + \theta\overline{y}_{i.} = x_{it}(\beta) + \theta[\overline{y}_{i.} - \overline{x}_{i.}(\beta)].$$

One can test H_o, by running the GAUSS-NEWTON regression on (2.4) evaluated at estimates that are root-*n* consistent under the null. In this case, *n* = *NT* and the restricted estimates would typically be non-linear least squares on (2.1). More specifically, the GNR amounts to regressing $y_{it} - Z_{it}\,(\beta,\theta)$ on the derivatives of the regression function $Z_{it}\,(\beta,\theta)$ with the respect to all the parameters, where both $Z_{it}\,(\beta,\theta)$ and its derivatives are evaluated under the null. Let $\tilde{\beta}$ denote the non-linear least squares estimate of β conditional on $\sigma^2_\mu = 0$. The derivatives are:

$$\partial Z_{it}\,(\beta,\theta)/\partial\beta = X_{it}(\beta) - \theta\overline{X}_{i.}\,(\beta) \tag{2.5}$$

$$\partial Z_{it}\,(\beta,\theta)/\partial\theta = \overline{y}_{i.} - \overline{x}_{i.}\,(\beta) \tag{2.6}$$

where

$$X_{it}(\beta) = \partial x_{it}(\beta)/\partial\beta, \tag{2.7}$$

$$\overline{X}_{i.}(\beta) = \partial\overline{x}_{i.}(\beta)/\partial\beta = [\sum_{t=1}^{T} \partial x_{it}/\partial\beta]/T = \sum_{t=1}^{T} X_{it}(\beta)/T.$$

Note that the derivative of θ with respect to σ^2_μ evaluated under H_o; $\sigma^2_\mu = 0$ is equal to $T/2\tilde{\sigma}^2$ where $\tilde{\sigma}^2 = \tilde{u}'\tilde{u}/n$ and $\tilde{u}$ denotes the vector of non-linear least squares with typical element $y_{it} - x_{it}(\tilde{\beta})$. The GNR becomes

$$(2.8) \qquad y_{it} - x_{it}(\tilde{\beta}) = X_{it}(\tilde{\beta})b + \frac{Tc}{2\tilde{\sigma}^2}[\overline{y}_{i.} - \overline{x}_{i.}(\tilde{\beta})] + \text{residual}.$$

This can be rewritten in vector form as

$$(2.9) \qquad \tilde{u} = \tilde{X}b + c_1\,TP\tilde{u} + \text{residuals}$$

where $c_1 = c/2\tilde{\sigma}^2$. $\tilde{X}$ denotes the nxK matrix of derivatives of the regression function $x_{it}(\beta)$ evaluated at $\tilde{\beta}$ and $P = (I_N \otimes J_T)/T$ is the averaging matrix over time. I_N is an identity matrix of dimension N, while J_T is a matrix of ones of dimension T. The typical element of $P\tilde{u} = \sum_{t=1}^{T} \tilde{u}_{it}/T$.

In case the original regression model in (2.1) is linear, $\tilde{X}$ becomes the matrix of regressors X, $\tilde{u}$ becomes the vector of OLS residuals and this GNR is simply the regression of OLS residuals on the matrix of regressors and the vector of OLS residuals summed over time:

$$(2.10) \qquad \tilde{u} = Xb + c_1TP\tilde{u} + \text{ residuals}$$

Subtracting $c_1\tilde{u}$ from both sides of (2.10) and dividing by $(1 - c_1)$ yields

$$(2.11) \qquad \tilde{u} = Xb_2 + c_2(TP - I_n)\tilde{u} + \text{ residuals}$$

where $b_2 = b/(1 - c_1)$ and $c_2 = c_1/(1 - c_1)$. It is then easy to show, see the Appendix, that $1/\sqrt{2}$ times the t-statistic for $c_2 = 0$ in (2.11) is asymptotically distributed as $N(0,1)$ under the null hypothesis.[1]

If the original regression model in (2.1) is linear, a simple test for H_o; $c_2 = 0$ can be derived as a variable addition test, *à la* PAGAN [1984] and PAGAN and HALL [1983], as follows:

$$(2.12) \qquad y = Xb_2 + c_2(TP - I_n)\tilde{u} + \text{ residuals}$$

This gives exactly the same sum of squared residuals and exactly the same t-statistic for $c_2 = 0$ as (2.11). Thus, for panel data linear regression models, a simple test for zero random individual effects amounts to running the original regression (2.1) with one additional regressor made up of the vector of OLS residuals summed over time and expressed as deviations from the original residuals. The test statistic is $1/\sqrt{2}$ times the t-statistic on the additional regressor. This is asymptotically distributed as $N(0,1)$ under the null hypothesis.

[1] I am grateful to an anonymous referee for this derivation.

3 Testing Linear and Log-Linear Error Components Regressions Against Box-Cox Alternatives

DAVIDSON and MACKINNON [1985] showed that the Double-Length Artificial regression (DLR) can be very useful in choosing between, and testing the specification of models that are linear or log-linear in the dependent variable. BALTAGI [1997a] extends this analysis to panel data regressions, where the choice between linear and log-linear models is complicated by the presence of error components. Following DAVIDSON and MACKINNON [1985], we consider the following two competing regression models:

$$(3.1) \quad y_{it} = \sum_{k=1}^{K} \beta_k X_{itk} + \sum_{s=1}^{S} \gamma_s Z_{its} + u_{it} \qquad i = 1,2,..,N; \; t = 1,2,..,T$$

$$(3.2) \qquad \log(y_{it}) = \sum_{k=1}^{K} \beta_k \log X_{itk} + \sum_{s=1}^{S} \gamma_s Z_{its} + u_{it}$$

where both X_{itk} and Z_{its} denote observations on independent variables with the distinction that the X_{itk} are subject to transformation, while the Z_{its} are not. For example, in estimating earnings from the panel study of income dynamics, y_{it} may denote the wage of the i-th individual at time t. X_{it} may include variables like years of full time work, experience and the number of weeks worked. These variables must take only positive values. Z_{it} may include dummy variables like union participation, marital status and time invariant variables like sex and race. We assume, as is typically the case that there is a constant term which is included in the Z_{its}'s. The disturbances follow a one-way error component model, given by (2.2).

Both (3.1) and (3.2) are special cases of the conventional BOX-COX model

$$(3.3) \qquad B(y_{it},\lambda) = \sum_{k=1}^{K} \beta_k B(X_{itk},\lambda) + \sum_{s=1}^{S} \gamma_s Z_{its} + u_{it}$$

where

$$(3.4) \qquad \begin{aligned} B(y_{it},\lambda) &= (y_{it}^{\lambda} - 1)/\lambda \quad \text{for } \lambda \neq 0 \\ &= \log(y_{it}) \quad \text{for } \lambda = 0 \end{aligned}$$

is the familiar BOX-COX transformation which requires that y_{it} is always positive. One could estimate (3.1), (3.2) and (3.3) by maximum likelihood and test the linear and log-linear specifications using a likelihood ratio test, see BOX and COX [1964]. The unrestricted model is (3.3) and for $\lambda = 1$ we get the linear model (3.1), while for $\lambda = 0$ we get the log-linear model (3.2). Estimating (3.3) can be difficult compared to estimating the restricted models

(3.1) and (3.2). This suggests a LAGRANGE Multiplier (LM) test which requires only the restricted maximum likelihood estimates. We focus on the LM test based on the DLR which is described more fully in DAVIDSON and MACKINNON [1984a, 1988].

Applying the FULLER and BATTESE [1974] transformation to (3.3) gives

$$B^*(y_{it},\lambda) = \sum_{k=1}^{K} \beta_k B^*(X_{itk},\lambda) + \sum_{s=1}^{S} \gamma_s Z^*_{its} + u^*_{it} \tag{3.5}$$

where $B^*(y_{it},\lambda) = [B(y_{it},\lambda) - \theta \sum_{t=1}^{T} B(y_{it},\lambda)/T]$. Similarly,

$$B^*(X_{itk},\lambda) = [B(X_{itk},\lambda) - \theta \sum_{t=1}^{T} B(X_{itk},\lambda)/T]$$

and $Z^*_{its} = Z_{its} - \theta \overline{Z}_{i.s}$

with $\overline{Z}_{i.s} = \sum_{t=1}^{T} Z_{its}/T$.

Also, $u^*_{it} = (u_{it} - \theta \overline{u}_{i.})$

with $\overline{u}_{i.} = \sum_{t=1}^{T} u_{it}/T$.

In this case, $\theta = 1 - (\sigma_\nu/\sigma_1)$ with $\sigma_1^2 = T\sigma_\mu^2 + \sigma_\nu^2$. The transformed disturbances u^*_{it} are NID$(0,\sigma_\nu^2)$.

In order to apply the DLR, we rewrite (3.5) as

$$f_{it}(y_{it},\varphi) = \in_{it} \qquad i = 1,..,N;\ t = 1,..,T \tag{3.6}$$

with $\in_{it} \sim$ NID(0,1). This means that

$$f_{it}(y_{it},\varphi) = [B^*(y_{it},\lambda) - \sum_{k=1}^{K} \beta_k B^*(X_{itk},\lambda) - \sum_{s=1}^{S} \gamma_s Z^*_{its}]/\sigma_\nu \tag{3.7}$$

and $\varphi = (\beta,\gamma,\lambda,\theta,\sigma_\nu)$. The contribution of the *it*-th observation to the loglikelihood is

$$\ell_{it}(y_{it},\varphi) = \text{const.} - (1/2)\, f^2_{it}(y_{it},\varphi) + J_{it}(y_{it},\varphi) \tag{3.8}$$

where $J_{it}(y_{it},\varphi) = \log|\partial f_{it}(y_{it},\varphi)/\partial y_{it}|$ is a Jacobian term. Defining

$$F_{itj}(y_{it},\varphi) = \partial f_{it}(y_{it},\varphi)/\partial \varphi_j$$

$$J_{itj}(y_{it},\varphi) = \partial J_{it}(y_{it},\varphi)/\partial \varphi_j$$

then $F(y,\varphi)$ and $J(y,\varphi)$ are the NT × $(K+S+3)$ matrices with typical elements $F_{itj}(y_{it},\varphi)$ and $J_{itj}(y_{it},\varphi)$. Let $f(y,\varphi)$ be the NT vector with typical elements $f_{it}(y,\varphi)$, then the DLR regression can be written as

$$\begin{bmatrix} f(y,\varphi) \\ \iota_{NT} \end{bmatrix} = \begin{bmatrix} -F(y,\varphi) \\ J(y,\varphi) \end{bmatrix} b + \text{ residuals} \tag{3.9}$$

where ι_{NT} denotes a vector of ones of dimension NT. This artificial regression is double-length with 2NT observations. Intuitively, Davidson and MacKinnon argue that each observation makes two contributions to the likelihood, one through the $(-1/2)\, f_{it}^2$ term and the other through the Jacobian term J_{it}. One of the properties of this DLR is that it generates the LM test in its score form as the explained sum of squares of (3.9) when the latter is evaluated at $\tilde{\varphi}$, the restricted maximum likelihood estimates of φ under the null. Since the total sum of squares of the restricted model is always 2NT, the LM statistic can be obtained as 2NT minus the restricted sum of squares residuals. This LM statistic is asymptotically distributed as χ^2 with degrees of freedom equal to the number of restrictions.

For the linear model, given in (3.1), the null hypothesis is $H_o^a, \lambda = 1$. In this case, (3.5) becomes

$$y_{it}^* = \sum_{k=1}^{K} \beta_k X_{itk}^* + \sum_{s=1}^{S} \gamma_s Z_{its}^* + u_{it}^* \tag{3.10}$$

where $y_{it}^* = (y_{it} - \theta \overline{y}_{i.})$, $X_{itk}^* = X_{itk} - \theta \overline{X}_{i.k}$ with $\overline{X}_{i.k} = \sum_{t=1}^{T} X_{itk}/T$. Following the derivations in Baltagi [1997a], the regressand of the DLR has a typical element $(\tilde{u}_{it}^*/\tilde{\sigma}_\nu)$ for the first NT artificial observations and 1 for the next NT observations. $\tilde{u}_{it}^*$ is the *it*-th residual from the restricted maximum likelihood estimator of the Fuller-Battese transformed linear model, given in (3.10), and $\tilde{\sigma}_\nu$ is the corresponding estimate of σ_ν. This restricted MLE can be obtained as iterative generalized least squares on (3.1) as described by Breusch [1987]. The typical elements for the first NT and the second NT observations of the regressors in (3.9) are given by:

for β_k : $[(X_{itk} - \tilde{\theta}\, \overline{X}_{i.k}) - (1 - \tilde{\theta})]/\tilde{\sigma}_\nu$ and 0;

for γ_s : $(Z_{its} - \tilde{\theta}\, \overline{Z}_{i.s})/\tilde{\sigma}_\nu$ and 0;

for σ_ν : $\tilde{u}_{it}^*/\tilde{\sigma}_\nu^2$ and $(-1/\tilde{\sigma}_\nu)$;

for θ : $[(\sum_{t=1}^{T} \tilde{u}_{it}/T) - (1 - \sum_{k=1}^{K} \tilde{\beta}_k)]/\tilde{\sigma}_\nu$ and $-1/(T - \tilde{\theta})$;

for λ : $\{\sum_{k=1}^{K} \beta_k [C(X_{itk},1) - \tilde{\theta} \sum_{t=1}^{T} C(X_{itk},1)/T] - [C(y_{it},1)$

$-\tilde{\theta} \sum_{t=1}^{T} C(y_{it},1)/T]\}/\tilde{\sigma}_\nu$ and $\log(y_{it})$;

where $C(y_{it},1) = y_{it} \log(y_{it}) - (y_{it} - 1)$ and $C(X_{itk},1) = X_{itk} \log(X_{itk}) - (X_{itk} - 1)$. The term $\tilde{u}_{it}$ refers to the residuals obtained from the linear model (3.1) by substituting the restricted MLE of (3.10). The test statistic is 2NT – the residuals sum of squares. This is asymptotically distributed as χ_1^2 under the null hypothesis H_o^a.

For the log-linear model given in (3.2), the null hypothesis is H_o^b; $\lambda = 0$. In this case, (3.5) becomes

$$\log^*(y_{it}) = \sum_{k=1}^{K} \beta_k \log^*(X_{itk}) + \sum_{s=1}^{S} \gamma_s Z_{its}^* + u_{it}^* \tag{3.11}$$

where

$$\log^*(y_{it}) = \log(y_{it}) - \theta \sum_{t=1}^{T} \log(y_{it})/T,$$

and

$$\log^*(X_{itk}) = \log(X_{itk}) - \theta \sum_{t=1}^{T} \log(X_{itk})/T,$$

with $Z_{its}^* = Z_{its} - \theta \overline{Z}_{i.s}$. Using the derivations in Baltagi [1997a], the regressand of the DLR has a typical element $(\widehat{u}_{it}^*/\widehat{\sigma}_\nu)$ for the first NT artificial observations and 1 for the next NT observations. $\widehat{u}_{it}^*$ the *it*-th residual from the restricted maximum likelihood estimator of the Fuller-Battese transformed log-linear model given in (3.11), and $\widehat{\sigma}_\nu$ is corresponding estimate of σ_ν. This restricted MLE can be obtained as iterative generalized least squares on (3.2) as described by Breusch [1987]. The typical elements for the first NT and the second NT observations of the regressors are given by:

for β_k : $[\log(X_{itk}) - \widehat{\theta} \sum_{t=1}^{T} \log(X_{itk}/T]/\widehat{\sigma}_\nu$ and 0;

for γ_s : $(Z_{its} - \widehat{\theta}\, \overline{Z}_{i.s})/\widehat{\sigma}_\nu$ and 0;

for σ_ν : $\widehat{u}_{it}^*/\widehat{\sigma}_\nu^2$ and $(-1/\widehat{\sigma}_\nu)$;

for θ : $(\sum_{t=1}^{T} \widehat{u}_{it}/T)/\widehat{\sigma}_\nu$ and $-1/(T - \widehat{\theta})$;

for λ : $\{\sum_{k=1}^{K} \widehat{\beta}_k [C(X_{itk},0) - \widehat{\theta} \sum_{t=1}^{T} C(X_{itk},0)/T] - C(y_{it},0)$

$-\widehat{\theta} \sum_{t=1}^{T} C(y_{it},0)/T]\}/\widehat{\sigma}_\nu$ and $\log(y_{it})$;

where

$$C(y_{it},0) = \lim_{\lambda \to 0} C(y_{it},\lambda) =$$

$$\lim_{\lambda \to 0} [\lambda y_{it}^\lambda \log(y_{it}) - (y_{it}^\lambda - 1)]/\lambda^2 = \{\log(y_{it})\}^2/2,$$

by L'HOPITAL's rule, see DAVIDSON and MACKINNON [1985]. Similarly, $C(X_{itk},0) = \{\log(X_{itk})\}^2/2$. The term $\widehat{u}_{it}$ refers to the residuals obtained from the log-linear model (3.2) by substituting the restricted MLE of (3.11). The test statistic is 2NT – the residuals sum of squares. This is asymptotically distributed as χ_1^2 under the null hypothesis H_o^b. BALTAGI [1997a] applied these DLR tests to a gasoline demand equation using a panel of 18 OECD countries over the period 1960-1978. In this case, both the linear and log-linear models were rejected in favor of a more general BOX-COX model.

This DLR can be easily extended to test jointly for functional form and random individual effects. More specifically, one can test H_o^c; $\lambda = 1$ and $\theta = 0$, that is, the model is linear with no random individual effects against the more general BOX-COX model given in (3.3) with disturbances following the random one-way error component model given in (2.2). Also, one can test H_o^d; $\lambda = 0$ and $\theta = 0$, that is, the model is log-linear in y and X with no random individual effects against the more general BOX-COX model described above. See the *Econometric Theory* problem by BALTAGI [1997b] and the solution by LI [1998].

Another application of the DLR to panel data is given by LARSON and WATTERS [1993]. This paper considers a BOX-COX model with disturbances that are allowed to be cross-sectionally heteroskedastic and time-wise auto-correlated. They apply a joint test of functional form and non-spherical disturbances to intrastate long distance demand for Southwestern Bell. The data set covers a five-state region observed quarterly over the period 1979-1988. Their results reject the logarithmic transformation on both the dependent and independent variables and is in favor of correcting for serial correlation and heteroskedasticity.

One limitation of the tests considered in this section is that the disturbances are assumed to be Normally distributed. Monte Carlo evidence by GODFREY, MCALEER and MCKENZIE [1988] show that the outer product gradient method for testing linear and logarithmic regression models does not perform well when the disturbances are non-Normal. This result may carry over to the DLR considered in this section under non-Normality of the disturbances.

4 Testing for Fixed Effects in Logit and Probit Models

Consider the fixed effects regression model

$$y_{it}^* = X_{it}'\beta + \mu_i + \nu_{it} \qquad i = 1,2,..,N;\ t = 1,2,..,T \tag{4.1}$$

where y_{it}^* denotes a (net) utility index derived by the i-th individual in the t-th time period from some action, like buying a car or participating in the labor force. X_{it} is a $Kx1$ vector of exogenous regressors, μ_i denotes the individual's fixed effects and $\nu_{it} \sim$ IID(0,1). We normalize the variance of ν_{it} to be unity since we only observe the sign of y^*_{it}. In fact, we let

$$y_{it} = 1 \text{ if } y^*_{it} > 0 \text{ and } y_{it} = 0 \text{ if } y^*_{it} \leqslant 0.$$

For this binary response model with fixed effects

$$\text{Pr}[y_{it} = 1] = \text{pr}[y^*_{it} > 0] = \text{Pr}[\nu_{it} > -X'_{it}\beta - \mu_i] = F(X'_{it}\beta + \mu_i)$$

where $F(.)$ denotes the cumulative function (c.d.f.) of ν_{it} which is symmetric around zero. This could be the Normal c.d.f. which is denoted by Φor the logistic c.d.f. which is denoted by Λ. It is important in panel data regression to test for H_o; $\mu_i = 0$ for $i = 1,..,N$. If H_o is not rejected, the estimation procedure is simple and utilizes the usual logit and probit procedures. However, if H_o is rejected, the maximum likelihood (ML) procedure is complicated by the presence of the fixed effects. For the logit model, CHAMBERLAIN [1980] suggested a conditional ML procedure which yields enormous computational simplifications. Unfortunately, this conditional ML approach does not yield similar computational simplifications for the fixed effects probit model.

BALTAGI [1995] suggested using the Binary Response Model Regression (BRMR) proposed by DAVIDSON and MACKINNON [1984b] to test for fixed effect in (4.1). In this case, the GNR ignoring heteroskedasticity of the error is given by

$$y_{it} - F(X'_{it}\beta + \mu_i) = f(X'_{it}\beta + \mu_i)X'_{it}b + f(X'_{it}\beta + \mu_i)Z'_{it}c + \text{ residuals}$$

where $f(.)$ denotes the probability density function (p.d.f.) of ν_{it}. Also, Z'_{it} is the *it*-th row of the NT $\times$ *N* matrix of individual dummies $Z = I_N \otimes \iota_T$ where I_N is an identity matrix of dimension *N* and ι_T is a vector of ones of dimension *T*. The heteroskedastic variance is given by

$$V_{it} = [F(X'_{it}\beta + \mu_i)][1 - F(X'_{it}\beta + \mu_i)]$$

and the GNR modified to handle this heteroskedasticity is given by:

$$\text{(4.2)} \qquad V_{it}^{-1/2}[y_{it} - F(X'_{it}\beta + \mu_i)] =$$

$$V_{it}^{-1/2} f(X'_{it}\beta + \mu_i)X'_{it}b + V_{it}^{-1/2} f(X'_{it}\beta + \mu_i)Z'_{it}c + \text{ residuals}$$

If we denote by $\tilde{\beta}$ the vector of ML estimates subject to the restriction H_o; $\mu_i = 0$, for $i = 1,2,..,N$; then one can test H_o by running the following BRMR:

$$\text{(4.3)} \qquad \tilde{V}_{it}^{-1/2}[y_{it} - F(X'_{it}\tilde{\beta})] =$$

$$\tilde{V}_{it}^{-1/2} f(X'_{it}\tilde{\beta})X'_{it}b + \tilde{V}_{it}^{-1/2} f(X'_{it}\tilde{\beta})Z'_{it}c + \text{ residuals}$$

where

$$\tilde{V}_{it} = [F(X'_{it}\tilde{\beta})][1 - F(X'_{it}\tilde{\beta})].$$

The test statistic in this case is the explained sum of squares from this BRMR. It will be asymptotically distributed as χ^2_N under the null hypothesis. Note that it will not be nR^2 since the total sum of squares is not equal to NT.

For the logit model, where $F(.)$ is $\Lambda(.)$, this BRMR simplifies to

$$\text{(4.4)} \qquad \tilde{\lambda}_{it}^{-1/2}[y_{it} - \Lambda(X'_{it}\tilde{\beta})] = \tilde{\lambda}_{it}^{1/2}X'_{it}b + \tilde{\lambda}_{it}^{1/2}Z'_{it}c + \text{ residuals}$$

where we made use of the fact that for the logistic distribution, its p.d.f. which is denoted by $\lambda(.)$ satisfies the following relationship with its c.d.f. $\Lambda(.)$:

$$\tilde{\lambda}_{it} = \lambda(X'_{it}\tilde{\beta}) = [\Lambda(X'_{it}\tilde{\beta})][1 - \Lambda(X'_{it}\tilde{\beta})].$$

In this case, the restricted ML estimates of β are simply the logit estimates ignoring the fixed effects. Therefore, this BRMR is simply a weighted least squares regression of logit residuals, ignoring the fixed effects, on the matrix of regressors X and the matrix of individual dummies Z. The weights are easily computable from the logistic p.d.f. $\tilde{\lambda}_{it}$. The regressand gets a $\tilde{\lambda}_{it}^{-1/2}$ weight, whereas the regressors get a $\tilde{\lambda}_{it}^{1/2}$ weight.
Similarly, for the probit model, where $F(.)$ is $\Phi(.)$, this BRMR simplifies to

$$\text{(4.5)} \quad \tilde{\Phi}_{it}^{-1/2}(1 - \tilde{\Phi}_{it})^{-1/2}[y_{it} - \tilde{\Phi}_{it}] = \tilde{\Phi}_{it}^{-1/2}(1 - \tilde{\Phi}_{it})^{-1/1}\tilde{\phi}_{it}X'_{it}b$$
$$+\tilde{\Phi}_{it}^{-1/2}(1 - \tilde{\Phi}_{it})^{-1/2}\tilde{\phi}_{it}Z'_{it}c + \text{ residuals}$$

where

$$\tilde{\Phi}_{it} = \Phi(X'_{it}\tilde{\beta}), \quad \tilde{\phi}_{it} = \phi(X'_{it}\tilde{\beta})$$

and $\phi(.)$ is the Normal p.d.f. In this case, the restricted ML estimates of β are simply the probit estimates ignoring the fixed effects. Therefore, this BRMR is simply a weighted least squares regression of probit residuals, ignoring the fixed effects, on the matrix of regressors X and the matrix of individual dummies Z. The weights are easily computable from the Normal c.d.f. Φ and its p.d.f. ϕ. See the *Econometric Theory* solution by Gurmu [1996].

5 Other Examples of Artificial Regressions in Panel Data

This paper surveyed three applications of the use of artificial regressions in panel data using the GNR, DLR and BRMR described in Davidson and MacKinnon [1993]. These applications are by no means the only ones. In fact, the panel data literature has some important examples of the use of artificial regressions for specification tests and I am sure that more can be derived. In this section we review some of the familiar tests.

The first is HAUSMAN's [1978] specification test. For the linear regression model

$$y_{it} = X'_{it}\beta + u_{it} \tag{5.1}$$

with error component disturbances given by (2.2), it is well known that HAUSMAN's test is based on the difference between the GLS and within estimators of the regression coefficients. In fact, HAUSMAN [1978] suggested its computation based on the following artificial regression:

$$y^*_{it} = X^{*\prime}_{it}b + \tilde{X}'_{it}c + \text{ residuals} \tag{5.2}$$

where $y^*_{it} = y_{it} - \theta\overline{y}_{i.}$ and $X^*_{it,k} = X_{it,k} - \theta\overline{X}_{i.,k}$ describes the FULLER and BATTESE [1974] transformation. While, $\tilde{X}_{it} = X_{it,k} - \overline{X}_{i.,k}$ describe the within transformation of the k-th regressor. HAUSMAN's test is based on testing $c = 0$ in (5.2). This is shown to be equivalent to testing that the contrast $q = \widehat{\beta}_{\text{GLS}} - \tilde{\beta}_{\text{Within}}$ is zero. In this case, $\widehat{\beta}_{\text{GLS}}$ denotes the GLS estimator of β obtained by regressing y^*_{it} on X^*_{it}, whereas $\tilde{\beta}_{\text{Within}}$ denotes the Within estimator obtained by regressing $\tilde{y}_{it} = y_{it} - \overline{y}_{i.}$ on $\tilde{X}_{it}$. Under the null hypothesis H_o; $E(\mu_i/X_{it}) = 0$, the GLS estimator is efficient while the within estimator is consistent. Under the alternative hypothesis that μ_i is correlated with X_{it}, the within estimator remains consistent while the GLS estimator is no longer consistent. This assumes that the disturbances in (2.2) are homoskedastic and not serially correlated. More recently, ARELLANO [1993] provided an alternative variable addition test to the HAUSMAN test which is robust to autocorrelation and heteroskedasticity of arbitrary form. In particular, ARELLANO [1993] suggests constructing the following regression:

$$\begin{pmatrix} y_i^+ \\ \overline{y}_i \end{pmatrix} = \begin{bmatrix} X_i^+ & 0 \\ \overline{X}'_i & \overline{X}'_i \end{bmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} u_i^+ \\ \overline{u}_i \end{pmatrix} \tag{5.3}$$

where $y_i^+ = (y^+_{i1},..,y^+_{iT})'$ and $X_i^+ = (X^+_{i1},..,X^+_{iT})'$ is a $T \times K$ matrix and $u_i^+ = (u^+_{i1},..,u^+_{iT})'$. Also

$$y^+_{it} = \left[\frac{T-t}{T-t+1}\right]^{1/2}\left[y_{it} - \frac{1}{(T-t)}(y_{i,t+1} + .. + y_{iT})\right] t = 1,2,..,T-1$$

$\overline{y}_i = \sum_{t=1}^{T} y_{it}/T$, X^+_{it}, $\overline{X}_i$, u^+_{it} and $\overline{u}_i$ are similarly defined. OLS on this model yields $\widehat{\beta} = \tilde{\beta}_{\text{Within}}$ and $\widehat{\gamma} = \widehat{\beta}_{\text{Between}} - \tilde{\beta}_{\text{Within}}$. HAUSMAN and TAYLOR [1981] showed that the WALD statistic for $\gamma = 0$ based on $\widehat{\gamma}$ yields a numerically identical test statistic to that based on $q = \widehat{\beta}_{\text{GLS}} - \tilde{\beta}_{\text{Within}}$. Therefore, Hausman's test can be obtained from the artificial regression (5.3) by testing for $\gamma = 0$. If the disturbances are heteroskedastic and/or serially correlated, then neither $\tilde{\beta}_{\text{Within}}$ nor $\widehat{\beta}_{\text{GLS}}$ are optimal under the null or alternative. Also, the standard formulae for the asymptotic variances of these estimators are no longer valid. Moreover, these estimators cannot be ranked in terms of efficiency so that the var(q) is not the difference of the two variances $\text{var}(\tilde{\beta}_W) - \text{var}(\widehat{\beta}_{\text{GLS}})$. Arellano [1993] suggests using White's [1984] robust variance-covariance matrix from OLS on (5.3) and applying a standard Wald Test for $\gamma = 0$ using these robust standard errors. This can be easily calculated using any standard regression package that computes White robust standard errors. This test is asymptotically distributed as χ^2_K under the null.

Modifying the set of additional variables in (5.3) so that the set of K additional regressors are replaced by KT additional regressors we get

$$\text{(5.4)} \qquad \begin{pmatrix} y_i^+ \\ \overline{y}_i \end{pmatrix} = \begin{pmatrix} X_i^+ & 0 \\ \overline{X}_i' & X_i' \end{pmatrix} \begin{pmatrix} \beta \\ \lambda \end{pmatrix} + \begin{pmatrix} u_i^+ \\ \overline{u}_i \end{pmatrix}$$

where $X_i = (X_{i1}', .., X_{iT}')'$ and λ is KT $\times$ 1. Chamberlain's [1982] test of correlated effects based on the reduced form approach turns out to be equivalent to testing for $\lambda = 0$ in (5.4). Once again this can be made robust to an arbitrary form of serial correlation and heteroskedasticity by using a Wald test for $\lambda = 0$ using White's [1984] robust standard errors. This test is asymptotically distributed as χ^2_{TK}. Note that this clarifies the relationship between the Hausman and Chamberlain tests. In fact, both tests can be computed as Wald tests from the artificial regressions in (5.3) and (5.4). Hausman's test can be considered as a special case of the Chamberlain test for $\lambda_1 = \lambda_2 = .. = \lambda_T = \gamma/T$. Arellano [1993] extends this analysis to dynamic models and to the case where some of the explanatory variables are known to be uncorrelated or weakly correlated with the individual effects.

Recently, Ahn and Low [1996] showed that Hausman's test statistic can be obtained from the artificial regression of the GLS residuals $(y_{it}^* - X_{it}^{*\prime}\widehat{\beta}_{\text{GLS}})$ on $\tilde{X}$ and $\overline{X}$, where $\tilde{X}$ has typical element $\tilde{X}_{it,k}$ and $\overline{X}$ is the matrix of regressors averaged over time. The test statistic is NT times the R^2 of this regression. Using (5.2), one can test H_o; $c = 0$ by running the GNR evaluated at the restricted estimators under the null. Knowing θ, the restricted estimates yield $b = \widehat{\beta}_{\text{GLS}}$ and $c = 0$. Therefore, the GNR on (5.2) regresses the GLS residuals $(y_{it}^* - X_{it}^{*\prime}\widehat{\beta}_{\text{GLS}})$ on the derivatives of the regression function with respect to b and c evaluated at $\widehat{\beta}_{\text{GLS}}$ and $c = 0$. These regressors are X_{it}^* and $\tilde{X}_{it}$, respectively. But X_{it}^* and $\tilde{X}_{it}$ span the same space as $\tilde{X}_{it}$ and $\overline{X}_{i.}$. This follows immediately from the definition of X_{it}^* and $\tilde{X}_{it}$ given above. Hence, this GNR yields the same regression sums of squares and therefore, the same Hausman test statistic as that proposed by Ahn and Low [1996], see the *Econometric Theory* problem of Baltagi [1997c].

Ahn and Low [1996] argue that Hausman's test can be generalized to test that each X_{it} is uncorrelated with μ_i and not simply that $\overline{X}_i$ is uncorrelated with μ_i. In this case, one computes NT times R^2 of the regression of GLS residuals $(y_{it}^* - X_{it}^{*\prime}\widehat{\beta}_{\mathrm{GLS}})$ on $\tilde{X}_{it}$ and $[X_{i1}',..,X_{iT}']$. This LM statistic is identical to Arellano's [1993] Wald statistic described earlier if the same estimates of the variance components are used. Ahn and Low [1996] argue that this test is recommended for testing the joint hypothesis of exogeneity of the regressors and the stability of the regression parameters over time. If the regression parameters are nonstationary over time, both $\widehat{\beta}_{\mathrm{GLS}}$ and $\tilde{\beta}_{\mathrm{Within}}$ are inconsistent even though the regressors are exogenous. Monte Carlo experiments were performed that showed that both the Hausman test and the Ahn and Low [1996] test have good power in detecting endogeneity of the regressors. However, the latter test dominates if the coefficients of the regressors are nonstationary. For Ahn and Low [1996], rejection of the null does not necessarily favor the within estimator since the latter estimator may be inconsistent. In this case, the authors recommend performing Chamberlain's [1982] test or the equivalent test proposed by Angrist and Newey [1991] which we consider next.

Chamberlain [1982] showed that the fixed effects specification imposes testable restrictions on the coefficients from regressions of all leads and lags of dependent variables on all leads and lags of independent variables. For the linear panel data model given in (5.1) with disturbances given by (2.2), Chamberlain [1982] specified the relationship between the unobserved individual effects and X_{it} as follows:

$$\mu_i = X_{i1}'\lambda_1 + .. + X_{iT}'\lambda_T + \in_i \tag{5.5}$$

where each λ_t is of dimension $K \times 1$ for $t = 1,2,..,T$. Let $y_i' = (y_{i1},..,y_{iT})$ and $X_i' = (X_{i1}',..,X_{iT}')$ and denote the "*reduced form*" regression of y_i' and X_i' by

$$y_i' = X_i'\pi + \eta_i \tag{5.6}$$

The restrictions between the reduced form and structural parameters are given by

$$\pi = (I_T \otimes \beta) + \lambda\iota_T' \tag{5.7}$$

with $\lambda = {}^\prime = (\lambda_1',..,\lambda_T')$. Chamberlain [1982] suggested estimation and testing be carried out using the minimum chi-square method where the mimimand is a χ^2 goodness of fit statistic for the restrictions on the reduced form. However, Angrist and Newey [1991] showed that this mimimand can be obtained as the sum of the T terms. Each term of this sum is simply the degrees of freedom times the R^2 from a regression of the within residuals for a particular period on all leads and lags of the independent variables. Angrist and Newey [1991] illustrate this test using two examples. The first example estimates and tests a number of models for the union-wage effect using five years of data from the National Longitudinal Survey of Youth (NLSY). They find that the assumption of fixed effects in an equation for union-wage effects is not at odds with the data. The second example considers a conventional human capital earnings function. They find that the fixed effects estimates of the return to schooling in the NLSY are roughtly twice those of ordinary least squares. However, the over-identification test suggest that the fixed effects assumption may be inappropriate for this model.

The third example is from WOOLDRIDGE [1995] who derives some simple variable addition tests for detecting sample selection bias in linear fixed effects panel data models. For the model given in (5.1) with disturbances given by (2.2), WOOLDRIDGE [1995] considers the fixed effects model where the μ_i's are correlated with X_{it}. However, the remainder disturbances ν_{it} are allowed to display arbitrary serial correlation and unconditional heteroskedasticity. The panel is unbalanced with the selection indicator vector for each individual i denoted by $s_i' = (s_{i1}, s_{i2}, .., s_{it})$. When $s_{it} = 1$, it is assumed that (X_{it}', y_{it}) is observed. The fixed effects estimator is given by

$$\tilde{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\tilde{X}_{it}\tilde{X}_{it}'\right)^{-1}\left(\sum_{i=t}^{N}\sum_{t=1}^{T} s_{it}\tilde{X}_{it}\tilde{y}_{it}\right) \tag{5.8}$$

where $\tilde{X}_{it}' = X_{it}' - \left(\sum_{r=1}^{T} s_{ir}X_{ir}'/T_i\right)$, $\tilde{y}_{it} = y_{it} - \left(\sum_{r=1}^{T} s_{ir}y_{ir}/T_i\right)$ and $T_i = \sum_{i=1}^{T} s_{it}$. A sufficient condition for the fixed estimator to be consistent and asymptotically Normal, as $N \to \infty$, is that $E(\nu_{it}/\mu_i, X_i', s_i') = 0$ for $t = 1,2,..,T$. Recall, that $X_i' = (X_{i1}', .., X_{iT}')$. Under this assumption, the selection process is strictly exogenous conditional on μ_i and X_i'.

WOOLDRIDGE [1995] considers two cases. The first is when the latent variable determining selection is partially observed. Define a latent variable

$$h_{it}^* = \delta_{to} + X_{i1}'\delta_{t1} + .. + X_{iT}'\delta_{tT} + \in_{it} \tag{5.9}$$

where $\in_{it}$ is independent of (μ_i, X_i'), δ_{tr} is a $Kx1$ vector of unknown parameters for $r = 1,2,..,T$ and $\in_{it} \sim N(0, \sigma_t^2)$.

The binary selection indicator is defined as $s_{it} = 1$ if $h_{it}^* > 0$. For this case, the censored variable $h_{it} = \max(0, h_{it}^*)$ is observed. For example, (5.1) could be a wage equation, and selection depends on whether or not individuals are working. If a person is working, the working hours h_{it} are recorded, and selection is determined by non-zero hours worked. This is what is meant by partial observability of the selection variable.

Because s_i is a function of $(X_i', \in_i')$ where $\in_i' = (\in_{i1}, .., \in_{iT})$, a sufficient condition for the fixed effects estimator to be consistent and asymptotically Normal as $N \to \infty$ is $E(\nu_{it}/\mu_i, X_i', \in_i') = 0$ for $t = 1,2,..,T$. The simplest alternative that imply selectivity bias is $E(\nu_{it}/\mu_i, X_i', \in_i') = E(\nu_{it}/\in_{it}) = \gamma \in_{it}$ for $t = 1,2,..,T$, with γ being an unknown scalar. Therefore,

$$E(y_{it}/\mu_i, X_i', \in_i', s_i') = E(y_{it}/\mu_i, X_i', \in_i') = \mu_i + X_{it}'\beta + \gamma \in_{it} \tag{5.10}$$

It follows that, if we could observe $\in_{it}$ when $s_{it} = 1$, then we could test for selectivity bias by including the $\in_{it}$ as an additional regressor in fixed effects estimation and testing H_o; $\gamma = 0$ using standard methods. While $\in_{it}$ cannot be observed, it can be estimated whenever $s_{it} = 1$ because $\in_{it}$ is simply the error of a Tobit model.

When h_{it} is observed, WOOLDRIDGE's [1995] test for selection bias is as follows:

Step 1: for each $t = 1,2,..,T$, estimate the equation

$$h_{it} = \max(0, X'_i\delta_t + \in_{it}) \tag{5.11}$$

by standard Tobit, where $\delta'_t = (\delta_{to}, \delta'_{t1}, .., \delta'_{tT})$ and X_i now has unity as its first element. For $s_{it} = 1$, let $\widehat{\in}_{it} = h_{it} - X'_i\widehat{\delta_t}$ denote the Tobit residuals.

Step 2: Estimate the equation

$$\tilde{y}_{it} = \tilde{X}'_{it}\beta + \gamma\tilde{\in}_{it} + \text{ residuals} \tag{5.12}$$

by pooled OLS using those observations for which $s_{it} = 1$. $\tilde{X}_{it}$ and $\tilde{y}_{it}$ were defined above, and

$$\tilde{\in}_{it} = \widehat{\in}_{it} - \left(\sum_{r=1}^{T} s_{ir}\widehat{\in}_{ir}/T\right). \tag{5.13}$$

Step 3: Test H_o; $\gamma = 0$ using the t-statistic for $\widehat{\gamma}$. A serial correlation and heteroskedasticity-robust standard error should be used unless $E[\nu'_i\nu_i/\mu_i, X'_i, s_i] = \sigma^2_\nu I_T$. This robust standard error is given in the Appendix to WOOLDRIDGE's paper.

The second case considered by WOOLDRIDGE is when h_{it} is not observed. In this case, one conditions on s_i rather than $\in_i$. Using iterated expectations, this gives

$$\begin{aligned} E(y_{it}/\mu_i, X'_i, s'_i) &= \mu_i + X'_{it}\beta + \gamma E(\in_{it}/\mu_i, X'_i, s'_i) \\ &= \mu_i + X'_{it}\beta + \gamma E(\in_{it}/X'_i, S'_i) \end{aligned} \tag{5.14}$$

If the $\in_{it}$ were independent across t, then $E(\in_{it}/X'_i, s'_i) = E(\in_{it}/X'_i, s_{it})$. The conditional expectation we need to estimate is $E[\in_{it}/X'_i, s_{it} = 1] = E[\in_{it}/X'_i, \in_{it} > -X'_i\delta_t]$. Assuming that the var($\in_{it}$) = 1, we get $E[\in_{it}/X'_i, \in_{it} > -X'_i\delta_t)] = \lambda(X'_i\delta_t)$ where $\lambda(.)$ denotes the inverse Mills ratio.

When h_{it} is not observed, WOOLDRIDGE's [1995] test for selection bias is as follows:

Step 1: For each $t = 1,2,..,T$, estimate the equation

$$\Pr(s_{it} = 1/X'_i] = \Phi(X'_i\delta_t) \tag{5.15}$$

using standard probit. For $s_{it} = 1$, compute $\widehat{\lambda}_{it} = \lambda(X'_i\widehat{\delta_t})$.

Step 2: Estimate the equation

$$\tilde{y}_{it} = \tilde{X}'_{it}\beta + \gamma\tilde{\lambda}_{it} + \text{ residuals} \tag{5.16}$$

by pooled OLS using those observations for which $s_{it} = 1$. $\tilde{X}_{it}$ and $\tilde{y}_{it}$ were defined above, and

$$\tilde{\lambda}_{it} = \widehat{\lambda}_{it} - \left(\sum_{r=1}^{T} s_{ir}\widehat{\lambda}_{ir}/T_i\right) \tag{5.17}$$

Step 3: Test H_o; $\gamma = 0$ using the *t*-statistic for $\gamma = 0$. Again, a serial correlation and heteroskedasticity-robust standard error is warranted unless

$$E(\nu'_i\nu_i/\mu_i, X'_i, s_i] = \sigma^2 I_T \quad \text{under } H_o.$$

Both tests proposed by WOOLDRIDGE [1995] are computationally simple involving variable addition tests. These require either Tobit residuals or inverse Mills ratios obtained from probit estimation for each time period. This is followed by fixed effects estimation.

For the random effects model, VERBEEK and NIJNAN [1992] suggest including three simple variables in the regression to check for the presence of selection bias. These are (i) the number of waves the *i*-th individual participates in the panel, T_i, (ii) a binary variable taking the value 1 if and only if the *i*-th individual is observed over the entire sample, $\prod_{r=1}^{T} s_{ir}$, and (iii) $s_{i,t-1}$ indicating whether the individual was present in the last period. Intuitively, testing the significance of these variables checks whether the pattern of missing observations affects the underlying regression. WOOLDRIDGE [1995] argues that the first two variables have no time variation and cannot be implemented in a fixed effects model. He suggested other variables to be used in place of $\widehat{\lambda}_{it}$ in a variable addition test during fixed effects estimation. These are $\sum_{r\neq t}^{T} s_{ir}$ and $\prod_{r\neq t}^{T} s_{ir}$. Such tests have the computational simplicity advantage and the need to only observe X_{it} when $s_{it} = 1$.

6 Conclusion

Specification testing is an integral part of econometrics, see PAGAN and HALL [1983], PAGAN [1984], GODFREY [1988] and WHITE [1994] to mention of few. It is also prominent in the panel data literature, see HSIAO [1986], and MÁTYÁS and SEVESTRE [1996]. This paper focuses on specification tests in panel data using artificial regressions. Some examples are given applying the GNR, DLR and BRMR to panel data regressions. In addition, several other artificial regression tests are surveyed including HAUSMAN's [1978] specification test, CHAMBERLAIN's [1982] omnibus goodness-of-fit test and WOOLDRIDGE's [1995] simple variable addition tests for selection bias. The important point to emphasize is that in many cases these artificial regressions provide the easiest way to compute specification tests, and in most cases provide a reasonably easy way to do so. For more details on the power and finite sample properties of the GNR, DLR and BRMR tests, see MACKINNON [1992] and DAVIDSON and MACKINNON [1993].

• References

AHN S.C., LOW S. (1996) – "A Reformulation of the Hausman Test for Regression Models with Pooled Cross-Section Time-Series Data", *Journal of Econometrics*, 71, pp. 309-319.

ANGRIST J.D., NEWEY W.K. (1991). – "Over-Identification Tests in Earnings Functions with Fixed Effects", *Journal of Business and Economic Statistics*, 9, pp. 317-323.

ARELLANO M. (1993). – "On the Testing of Correlated Effects with Panel Data", *Journal of Econometrics*, 59, pp. 87-97.

BALTAGI B.H. (1995). – "Testing for Fixed Effects in Logit and Probit Models Using and Artificial Regression", Problem 95.5.4, *Econometric Theory*, 11, pp. 1179.

BALTAGI B.H. (1996). – "Testing for Random Individual and Time Effects Using a GAUSS-NEWTON Regression", *Economics Letters*, 50, pp. 189-192.

BALTAGI B.H. (1997a). – "Testing Linear and Loglinear Error Components Regressions against BOX-COX Alternatives, *Statistics and Probability Letters*, 33, pp. 63-68.

BALTAGI B.H. (1997b). – "A Joint Test for Functional Form and Random Individual Effects", Problem 97.1.3, *Econometric Theory*, 13, pp. 307-308.

BALTAGI B.H. (1997c). – "HAUSMAN's Specification Test as a GAUSS-NEWTON Regression", Problem 97.4.1, *Econometric Theory*, 13, pp. 463.

BOX G.E.P., COX D.R. (1964). – "An Analysis of Transformations", *Journal of the Royal Statistical Society*, Series B, 26, pp. 211-252.

BREUSCH T.S. (1987). – "Maximum Likelihood Estimation of Random Effects Models", *Journal of Econometrics*, 36, pp. 383-389.

BREUSCH T.S., PAGAN A.R. (1980). – "The LAGRANGE Multiplier Test and its Applications to Model Specification in Econometrics", *Review of Economic Studies*, 47, pp. 239-253.

CHAMBERLAIN G. (1980). – "Analysis of Covariance with Qualitative Data", *Review of Economic Studies*, 47, pp. 225-238.

CHAMBERLAIN G. (1982). – "Multivariate Regression Models of Panel Data", *Journal of Econometrics*, 18, pp. 5-46.

DAVIDSON R., MACKINNON J.G. (1984a). – "Model Specification Tests Based on Artificial Linear Regressions", *International Economic Review*, 25, pp. 485-502.

DAVIDSON R., MACKINNON J.G. (1984b). – "Convenient Specification Tests for Logit and Probit Models", *Journal of Econometrics*, 25, pp. 241-262.

DAVIDSON R., MACKINNON J.G. (1985). – "Testing Linear and Log-Linear Regressions against BOX-COX Alternatives, *Canadian Journal of Economics*, 18, pp. 499-517.

DAVIDSON R., MACKINNON J.G. (1988). – "Double-Length Artificial Regressions", *Oxford Bulletin of Economics and Statistics*, 50, pp. 203-217.

DAVIDSON R., MACKINNON J.G. (1993). – *Estimation and Inference in Econometrics*, Oxford University Press: New York.

FULLER W.A., BATTESE G.E. (1974). – "Estimation of Linear Models with Cross-Error Structure", *Journal of Econometrics*, 2, pp. 67-78.

GODFREY L.G. (1988). – *Misspecification Tests in Econometrics: the LAGRANGE Multiplier Principle and Other Approaches*, Cambridge University Press: Cambridge.

GODFREY L.G., MCALEER M., MCKENZIE C.R. (1988). – “Variable Addition and LAGRANGE Multiplier Tests for Linear and Logarithmic Regression Models”, *Review of Economics and Statistics*, 70, pp. 492-503.

GURMU S. (1996). – “Testing for Fixed Effects in Logit and Probit Models Using an Artificial Regression”, Solution 95.5.4, *Econometric Theory*, 12, pp. 872-874.

HAUSMAN J.A. (1978). – “Specification Tests in Econometrics”, *Econometrica*, 46, pp. 1251-1271.

HAUSMAN J.A., TAYLOR W.E. (1981). – “Panel Data and Unobservable Individual Effects”, *Econometrica*, 49, pp. 1377-1398.

HONDA Y. (1985). – “Testing the Error Components Model with Non-Normal Disturbances”, *Review of Economic Studies*, 52, pp. 681-690.

HSIAO C. (1986). – *Analysis of Panel Data*, Cambridge University Press: Cambridge.

LARSON A.C., WATTERS J.S. (1993). – “A Convenient Test of Functional Form for Pooled Econometric Models”, *Empirical Economics*, 18, pp. 271-280.

LI D. (1998). – “A Joint Test for Functional Form and Random Individual Effects”, Solution 97.1.3, *Econometric Theory*, 14, pp. 154-156.

MÁTYÁS L., SEVESTRE P. (1996). – *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, Kluwer Academic Publishers: Dordrecht.

MACKINNON J.G. (1992). – “Model Specification Tests and Artificial Regressions”, *Journal of Economic Literature*, 30, pp. 102-146.

PAGAN A.R. (1984). – “Model Evaluation by Variable Addition”, in D.F. Hendry and K.F. Wallis, eds., *Econometrics and Quantitative Economics*, Basil Blackwell: Oxford, pp. 103-133.

PAGAN A.R., HALL A.D. (1983). – “Diagnostic Tests and Residual Analysis”, *Econometric Reviews*, 2, pp. 159-218.

VERBEEK M., NIJMAN T. (1992). – “Testing for Selectivity bias in Panel Data Models”, *International Economic Review*, 33, pp. 681-703.

WHITE H. (1984). – *Asymptotic Theory for Econometrics*, Academic Press: New York.

WHITE H. (1994). – *Estimation, Inference and Specification Analysis*, Cambridge University Press: Cambridge.

WOOLDRIDGE J.M. (1995). – “Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions”, *Journal of Econometrics*, 68, pp. 115-132.

APPENDIX

This Appendix shows the relationship between the Breusch and Pagan [1980] LM test for H_o; $\sigma_\mu^2 = 0$ and the GNR test-statistic derived in Section 2.

The Breusch and Pagan LM statistic for H_o; $\sigma_\mu^2 = 0$ is given by

$$\text{(A.1)} \qquad \text{LM} = \frac{NT}{2(T-1)}\left[\frac{T\tilde{u}'P\tilde{u}}{\tilde{u}'\tilde{u}}\right]^2 = \frac{n}{2(T-1)}(T\lambda_{\text{LM}} - 1)^2$$

where $n = NT$, $\tilde{u}$ denote the OLS residuals of the linear regression model version of (2.1) and $\lambda_{\text{LM}} = \tilde{u}'P\tilde{u}/\tilde{u}'\tilde{u}$. This LM statistic is asymptotically distributed as χ_1^2 under the null hypothesis. For the one-sided null hypothesis, Honda [1985] showed that $\sqrt{\frac{n}{2(T-1)}}(T\lambda_{\text{LM}}-1)$ is asymptotically distributed as $N(0,1)$.

Using the Frisch-Waugh and Lovell Theorem, see Davidson and MacKinnon [1993], the OLS estimates and residuals from (2.11) can be obtained from (2.12) after premultiplying it by $M = I_n - P_X$ where $P_X = X(X'X)^{-1}X'$, *i.e.*,

$$\text{(A.2)} \qquad \tilde{u} = c_2 M(TP - I_n)\tilde{u} + \text{ residuals}$$

since $MX = 0$. The OLS estimate of c_2 is

$$\text{(A.3)} \qquad \widehat{c}_2 = \tilde{u}'(TP - I_n)\tilde{u}/\tilde{u}'(TP - I_n)M(TP - I_n)\tilde{u}$$

since *M* is idempotent. Dividing up and down by $\tilde{u}'\tilde{u}$ yields

$$\text{(A.4)} \qquad \widehat{c}_2 = (T\lambda_{\text{LM}} - 1)/d$$

where

$$\text{(A.5)} \quad \text{d} = T^2\lambda_{\text{LM}} - 2T\lambda_{\text{LM}} + 1 - [\tilde{u}'(TP - I_n)P_X(TP - I_n)\tilde{u}/\tilde{u}'\tilde{u}]$$

It is easy to show that $\text{plim}\, T\lambda_{\text{LM}} = 1$. Hence,

$$\text{(A.6)} \qquad \text{plim}\frac{1}{T-1}d = 1$$

since the last term of *d* tends to zero. Therefore, $\widehat{c}_2$ is superconsistent under the null with $\text{plim}\,(T-1)\widehat{c}_2 = 0$. Also,

$$\sqrt{\frac{n(T-1)}{2}}\widehat{c}_2 = \left(\frac{1}{T-1}d\right)^{-1}\sqrt{\frac{n}{2(T-1)}}(T\lambda_{\text{LM}} - 1)$$

wich is asymptotically distributed as $N(0,1)$ using HONDA's [1985] result and (A.6). Now consider the *t*-statistic for $c_2 = 0$ in (2.11). This yields

$$\begin{aligned} t &= \widehat{c}_2/\text{s.e.}(\widehat{c}_2) = d^{-1}(T\lambda_{\text{LM}} - 1)/\sqrt{\sigma^2/\tilde{u}'\tilde{u}d} \\ \text{(A.7)} \qquad &= \sqrt{n}\left(\frac{\tilde{u}'\tilde{u}/n}{\sigma^2}\right)^{1/2}(T-1)^{-1/2}\left(\frac{1}{T-1}d\right)^{-1/2}(T\lambda_{\text{LM}}-1) \\ &\overset{a}{=} \sqrt{\frac{n}{T-1}}(T\lambda_{\text{LM}}-1) \end{aligned}$$

and this asymptotically distributed as $N(0,2)$ under the null. Therefore $t/\sqrt{2}$ is asymptotically distributed as $N(0,1)$ and $t^2/2$ as χ_1^2 under the null. It is also easy to show that t^2 is n times the uncentered R^2 of the regression in (2.11). Hence, $nR_u^2/2$ is also distributed as χ_1^2 under the null.