Unobservable Factors and Panel Data Sets: an Investigation in the Labour Market

Eleonora Patacchini
University of Southampton

Abstract

This paper investigates the effects of unobservable factors that, as is well-known, contaminate two of the variables most used in labour market research, namely the stock of unemployment and the stock of vacancies. Using a matching function framework, we compare different panel data estimators using a number of appropriate Hausman tests robust to deviations from the classical errors assumptions. The relevance of the choice of the model specification is underlined. It is shown to what extent conclusions lacking a rigorous statistical analysis may be misleading.

Keywords: panel data models, hypothesis testing, measurement errors, unobserved heterogeneity, unemployment and vacancy data

JEL Classification: C12, C13, C23, J60, J64

1 Introduction

It is well-known that measurement errors are extremely relevant in data collection. Even though the problem has given rise to a certain amount of theoretical interest, most applied econometric studies do not address this issue. In the analysis of uni-dimensional data sets, i.e. time series or cross sections, classical errors-in-variables models have not been applied widely mainly because it is often not possible to find valid instrumental variables among the variables included in those models. External variables are required in order to identify the structural parameters of interest. Furthermore, different kinds of measurement errors and other unobservable factors may affect our data. Some assumptions of the classical errors-in-variables model are often not sustainable in many empirical cases. When a panel data set is at hand it may be possible to handle these issues, since instrumental variables can be found within the model. Moreover, pooling cross sectional and time series observations, the econometrics of panel data offers a variety of different estimators for the same parameter, and the behavior of such estimators in the presence of unobserved factors affecting the data may be analyzed. Therefore, it is possible to acquire some knowledge about the kind of errors of specification involved by checking whether they can actually account for the sign and order of magnitude of the observed discrepancies between estimators. Pursuing this approach, we present the methodological development of a panel data econometric procedure aiming to offer a preliminary analysis of the data set in order to check for the presence of relevant sources of bias hidden in the data. As the presence of such unobserved factors may invalidate the estimation results, it is essential to use suitable estimators when different sources of bias are discovered. Our procedure checks for the presence of different "unobservables" and indicates which estimators are likely to give the most reliable results for the analysis of a certain data set.

Specifically, starting from a static model, we formally compare different panel data estimates of the same parameter using a number of appropriate Hausman tests robust to deviations from the classical errors assumption, i.e. the so called HR-test. This technique is based on the construction of the Hausman statistic using an artificial regression, as proposed by Arellano (1993). An auxiliary regression can be a calculating device that allow us to estimate a suitably constructed covariance matrix. Hence, it can be used us to estimate covariances matrices between estimators that cannot be ranked in terms of efficiency. Furthermore, the application of White's formulae (White, 1984) in the panel data case will lead to heteroskedasticity and autocorrelation consistent estimators. Therefore, an artificial regression can be used to construct a test for the comparison for a number of different estimators which is robust to deviations from the assumption of spherical disturbances.[1]

The paper is organized as follows.

[1] For a methodological revision of the use of the Hausman test for testing correlated effects with panel data, a derivation of the results of an application of White's formulae for estimators of covariances matrices (White, 1984) in a panel data context and a formal implementation of the HR-test, see Patacchini (2001).

Section 2 illustrates the above citated econometric procedure, that can be considered as a guide towards the choice of the most reliable model specification. It can be applied to every longitudinal data set.

Section 3 presents an empirical application of the methodology to a longitudinal data set of 277 travel-to-work areas (TTWAs) in the UK, observed monthly for the period 1996 - 2000. The primary aim is to analyze the effects of unobservable factors that, as is well-known, affect two of the variables most used in labour market research, namely the stock of unemployment and the stock of vacancies. A matching function framework is used. The relevance of the model specification is emphasized and it is shown to what extent conclusions based only on a visual inspection of the estimation results may be misleading.

2 A Diagnostic Analysis on a Panel Data Set: an Econometric Methodology

The Hausman test (Hausman, 1978) is the standard procedure used in empirical work in order to discriminate between different estimators. In panel data modelling, it is widely used as a test for correlated effects, i.e. to investigate the presence of unobserved heterogeneity across units correlated with the explanatory variables (Hausman and Taylor, 1981). The two estimators involved in the implementation of the test are the Within Group and the Balestra-Nerlove estimator; both OLS estimators constructed on different transformations of the data. The consistency of the two estimators, a requirement for the Hausman test to work properly, is almost never questioned in empirical studies. However, if we are in presence of measurement errors-in-variables least square estimators not only lose their efficiency but also their consistency. We may end up comparing two inconsistent estimators.

Moreover, measurement errors can have different impact using different transformations of the data. For instance, if we use first differences the bias can be magnified (Griliches and Hausman, 1986). As a consequence, the probability limits of two estimators calculated on different transformations of the data may be different. Thus, in the presence of strong measurement errors-in-variables OLS on the model in levels and OLS on the model in deviations (or first differences) would turn out to be different regardless of whether unobserved heterogeneity really matters. We may end up attributing the bias of our results to unobservable individual characteristics while it could be that the measurement errors are playing a major role. Consequently the specification of the model adopted could be inappropriate. In such contexts the use of the standard Hausman test is not methodologically correct and it may lead to unreliable results.

The econometrics of panel data, offering a variety of different estimators for the same parameter, can help us to deal with this issue. The failure of the assumption of consistency are often related to unobservable factors difficult to detect and to treat properly. The structure of a panel data set can be useful to distinguish among different sources of bias and can allow us to control for the effects of different kind of unobservable factors. Using the "repeated measurement property" of a panel data set, i.e. each cross sectional observation is followed over time, we can construct different kinds of instrumental variables inside the data set. Assuming a specific structure of the measurement errors, we can find instrumental variables estimators that remain consistent. Hence it is still possible to use the Hausman test framework but the two estimators we compare have to be chosen in an appropriate way.

We present an econometric procedure aiming to distinguish the effects of different kinds of "unobservables" on the estimators of the parameters in a panel data model in order to choose the most reliable specification.

The outline of the procedure is as follows (see Figure 1).

At a first stage, the Within Group estimator (OLS on the model in deviations from the individual time-means), which controls only for unobserved heterogeneity bias, is compared with a Generalized Instrumental Variables estimator on the model in deviations from the individual time-means, which controls for both measurement error bias and unobserved heterogeneity bias. A significant difference in the two estimators gives evidence of measurement problems in the data.

In this case, we investigate if unobserved individual characteristics matter also, by comparing the Generalized Instrumental Variables estimator on the model in levels, which controls only for measurement error bias with the Generalized Instrumental Variables estimator on the model in deviations from the individual time-means which controls for both measurement error bias and unobserved heterogeneity bias. If we find a significant difference in these two estimators we can infer that unobserved heterogeneity is also an important potential source of bias. On the contrary, if this difference is not significant we can conclude that the most important issue to control for is measurement problems in the data set.

On the other hand, if the test performed at the first step gives us insignificant results, we can conclude that measurement bias is not a major issue and we continue our diagnostic procedure comparing OLS on the model in levels and OLS on the model in deviations. The OLS estimator on the model in levels does not control for any kind of bias while the OLS on the model in deviations, i.e. the Within Group estimator, rules out the heterogeneity bias. A significant difference in the two estimators gives us evidence of unobserved heterogeneity bias in our data set.

It is worth noting how much the sequence of these tests matters. If we compare at the first step, as is common in empirical work, OLS on the model in levels (or the Between Group estimator) and OLS on the model in deviations (i.e. the Within Group estimator) we cannot distinguish what is the source of the bias because measurement errors have different effects in models in levels and in deviations from the mean, as previously emphasized.

We set out formally the comparisons using a number of appropriate HR-tests, i.e. panel data versions of the Hausman test robust to deviations from the classical errors assumptions. The hypothesis underlying the construction of the Hausman statistic (Hausman, 1978) are often too strong in most of the empirical cases. Although it is a well-know problem in empirical work, a robustified version of the Hausman Test is not directly implemented in any standard econometric software.

In what follows, we formalize the above outlined methodology in a way that can be implemented in most of the econometric packages.

We construct artificial augmented variables that are transformed as required by each model and insert them in a suitable matrix form, so that running an estimation method on these auxiliary regressions we end up estimating directly the difference of the two estimators we are interested in. Using the White's HAC estimators for the variances and performing an appropriate Wald test, we test exactly the equality of the two estimators without any assumptions on the disturbances.[2]

2.1 IVD versus WG: The HR-Test

The first step of the diagnostic procedure requires us to compare a Generalized Instrumental Variables estimator on data in deviations from individual time-means (or differences), hereafter IVD estimator, with the Within Group estimator, OLS estimator in data on deviations from individual time-means, hereafter WG estimator, in order to investigate the importance of measurement errors-in-variables, hereafter ME.

Particular care is required in the choice of the instruments we use. In order to apply the Hausman framework, we have to compare two estimators that are both consistent under the null hypothesis (one more efficient) and one consistent and the other inconsistent under the alternative. If the null hypothesis of no ME is satisfied, the WG estimator is more efficient than an IVD estimator but the instruments has to be chosen in a way such that the consistency of the IVD estimator has to hold when the null hypothesis is violated.

ME may arise under different forms, each of them having different effects on the estimators that are used. It is not possible to construct a reliable test for the presence of arbitrary ME. Panel data sets can help us with this issue because they provide a variety of different types of instrumental variables. However, the choice of the instruments has to be related to a specific structure of the measurement error in order to guarantee their validity.

Suppose, for instance, that we want to test for the presence of ME that may have a period specific component, as may arise in empirical work.

Consider the errors-in-variables panel data model

$$y_{it} = \beta x^*_{it} + \varepsilon_{it}; \quad i = 1; \ldots; N; \quad T = 1; \ldots; T; \tag{1}$$

where

$$x^*_{it} = x_{it} + m_{it}; \tag{2}$$

$$m_{it} = \mu_t + \xi_{it}; \tag{3}$$

$$x^*_{it};\ \varepsilon_{jt} \quad \text{independent for all } t \text{ and } i \neq j: \tag{4}$$

[2] The Stata 7 routines, that have been written for the empirical application of the methodology presented in Section 3 are available on request.

The process of the measurement error, i.e. m_{it}; consists of a time-specific effect, i.e. θ_t; and of a white noise component, i.e. ξ_{it}.

Substituting, we obtain

$$\begin{aligned} y_{it} &= \beta x_{it} + \delta_t + e_{it}; \\ \delta_t &= -\beta \theta_t; \\ e_{it} &= -\beta \xi_{it} + \varepsilon_{it}. \end{aligned}$$

If we define v_{it}, the new composite disturbance component, as

$$v_{it} = \delta_t + e_{it};$$

the basic assumption for the consistency of the OLS estimators, i.e. $E(v_{it} \mid x_{it}) = 0$; does not hold any longer and it is now replaced with

$$E(v_{it} \mid x_{it}) = \delta_t; \qquad t = 1; \ldots; T.$$

Moreover, the problem remains when we transform the model in deviations from the individual time-means.

Even when $T \to \infty$,

$$E(v_{it} - \overline{v_{i\cdot}} \mid x_{it} - \overline{x_{i\cdot}}) = \delta_t;$$

where

$$\overline{v_{i\cdot}} = \sum_{t=1}^{T} v_{it}; \quad \overline{x_{i\cdot}} = \sum_{t=1}^{T} x_{it}.$$

Hence, the WG estimator will be inconsistent.

However, by virtue of assumption (4)

$$E(v_{it} - \overline{v_{i\cdot}} \mid x_{jt} - \overline{x_{j\cdot}}) = 0; \qquad i \neq j.$$

Therefore, if we use as an instrument for the within variation of individual i, the within variation of individual j, we obtain an IVD estimator which is consistent in presence of measurement error having the specified structure. Thus, a HR-test for the comparison between the WG estimator and the above constructed IVD estimator can be applied and provides reliable results about the presence of measurement errors with a time-specific component.

It is worth noting that it is not possible to distinguish the effects of a measurement error with an individual-specific component from the ones arising from the presence of correlated effects. In the procedure we propose, this issue is investigated in a second step. If in this further stage we state the importance of unobserved heterogeneity bias, we can use the results of the first step to choose between a two-way and one-way panel data model. Specifically, a rejection of the test at the second stage means that fixed effects may be strong. A rejection of the test at the first stage means that measurement errors with a period specific component may be an important issue. The combination of these results leads us to the choice of a two-way panel data specification, i.e. $y_{it} = \beta x_{it} + \delta_t + e_{it}$, instead of the one-way, i.e. $y_{it} = \beta x_{it} + e_{it}$.

Another case, that can be worth investigating in the first step, is the presence of measurement errors that follow a moving average or autoregressive process. In this context, the instrumental variables have to be chosen according to the structure of the dynamic process. For instance, if we want to test for measurement errors following a moving average process of order one (MA(1) ME), possible valid instruments for a variable at time t are all the lag values of the same variable up to lag $(t - 2)$.

From the discussion above, it is clear that in the first step of the diagnostic procedure it is important to construct a number of artificial regressions, one for each structure of the measurement errors to be tested. The difference between them is only in the appropriate choice of the instruments. The structure of the methodology is the same.

In what follows, we will explain in detail how to formulate in practice the HR-test for the comparison between the WG estimator and an arbitrary IVD estimator, using a standard econometric package. There is not a straightforward way. Unlike the artificial regression proposed by Arellano (1993) and other cases we will analyse later on, here we do not directly compare the same estimation method applied on different transformations of the data. In other words, it is not just necessary to manipulate the data according to different transformations, insert the new variables in a auxiliary regression and then run an estimation method. Some preliminaries are needed.

In static models, the most efficient Generalized Instrumental Variables estimator is obtained by projecting the variables to be instrumented in the space generated by the instruments. This is a case where the instruments are orthogonal to the initial errors and especially correlated with the initial regressors. It can be shown that, given the properties of the projection matrix, it is equivalent to run OLS in a regression where the regressors are the projected variables.[3]

Consider a general model in matrix notation

$$Y = X\beta + U.$$

First transform the data according to the within transformation, i.e. deviations from the mean

$$Y^* = X^*\beta + U^*.$$

Then choose the instrumental matrix according to the structure of the measurement errors we want to test for, say Z. Project the variables we want to instrument in the space generated by Z

$$\widehat{X}^* = P_Z X^*$$

[3] For further details and an extensive discussion on these issues see Bowden and Turkington (1984).

where

$$P_Z = Z(Z'Z)^{-1}Z'.$$

Regress Y^* on $\tilde{X}^*$[4]

$$\hat{\beta}_{ivd} = (\tilde{X}^{*\prime}\tilde{X}^*)^{-1}\tilde{X}^*Y^*.$$

For the single individual i, construct the system

$$\left\{ \begin{array}{l} \tilde{y}_i^* = \tilde{x}_i^*\beta + \tilde{\varepsilon}_i^* \\ y_i^* = x_i^*\beta + \varepsilon_i^*. \end{array} \right.$$

Estimating by OLS the first group of equations, i.e. the ones in levels, we obtain the IVD estimator, i.e. $\hat{\beta}_{ivd}$. Estimating by OLS the second group, i.e. equations in deviations, we obtain the WG estimator, i.e. $\hat{\beta}_{wg}$.

Let

$$\beta_{ivd} = E\left(\hat{\beta}_{ivd}\right)$$

and

$$\beta_{wg} = E\left(\hat{\beta}_{wg}\right).$$

Rewrite the system as

$$\left\{ \begin{array}{l} \tilde{y}_i^* = \tilde{x}_i^*\beta_{ivd} + \tilde{\varepsilon}_i^* - \tilde{x}_i^*\beta_{wg} + \tilde{x}_i^*\beta_{wg} \\ y_i^* = x_i^*\beta_{wg} + \varepsilon_i^*. \end{array} \right.$$

Rearranging, we obtain

$$\left\{ \begin{array}{l} \tilde{y}_i^* = \tilde{x}_i^*\left(\beta_{ivd} - \beta_{wg}\right) + \tilde{x}_i^*\beta_{wg} + \tilde{\varepsilon}_i^* \\ y_i^* = x_i^*\beta_{wg} + \varepsilon_i^*. \end{array} \right.$$

Call

$$Y_i^+ = \begin{pmatrix} \tilde{y}_i^* \\ y_i^* \end{pmatrix}; \quad W_i^+ = \begin{pmatrix} \tilde{x}_i^* & \tilde{x}_i^* \\ 0 & x_i^* \end{pmatrix};$$

$$\beta^+ = \begin{pmatrix} \beta_{ivd} - \beta_{wg} \\ \beta_{wg} \end{pmatrix}; \quad \varepsilon_i^+ = \begin{pmatrix} \tilde{\varepsilon}_i^* \\ \varepsilon_i^* \end{pmatrix}.$$

[4] Recall that this is only a different reformulation of the IV estimators because the projection matrix is idempotent, i.e. $P_Z' = P_Z$ and $P_Z'P_Z = P_Z$;

$$\begin{aligned} \hat{\beta}_{ivd} &= (X^{*\prime}P_Z X^*)^{-1}X^{*\prime}P_Z Y^* \\ &= (X^{*\prime}P_Z' P_Z X^*)^{-1}X^{*\prime}P_Z' P_Z Y^* \\ &= (\tilde{X}^{*\prime}\tilde{X}^*)^{-1}\tilde{X}^{*\prime}Y^*. \end{aligned}$$

The augmented auxiliary model would be

$$Y_i^+ = W_i^+ \beta^+ + \varepsilon_i^+; \qquad i = 1;\ldots;N:$$

If we estimate β^+ with OLS, then we estimate the covariance matrix of β^+ using White's formulae and we perform an appropriate Wald test, we obtain a reliable HR-test comparing the two estimators we are interested in, namely $\hat{\beta}_{ivd}$ and $\hat{\beta}_{wg}$. In this context this procedure provides a reliable panel data implementation of the Hausman test for the presence of ME.

2.2 OLSL versus WG: The HR-Test

If the results of the test at the first stage provide evidence that ME can be neglected, the widespread practice to test for correlated effects using the comparison between OLS in levels, hereafter OLSL and in deviations, i.e. the WG estimator, is correct. The Hausman and Taylor (1981) set-up can be applied. However it is recommended to use a robustified version in order to control for the possible presence of non spherical disturbances.

The HR-test can be set out as follows. We first construct the auxiliary regression that was proposed by Arellano (1993) to test for random versus fixed effects in a static panel data model.

Consider the general panel data model for individual i

$$\underset{(T\times 1)}{Y_i} = \underset{(T\times K)}{X_i}\beta + \underset{(T\times 1)}{V_i}; \qquad i = 1;\ldots;N:$$

This system of T equations in levels can be transformed into $(T-1)$ equations in orthogonal deviations and one in averages. We obtain

$$\begin{cases} y_i^* = x_i^*\beta + \varepsilon_i^* & \rightarrow \ (T-1) \text{ equations} \\ \overline{y_i} = \overline{x_i}\beta + \overline{\varepsilon_i} & \rightarrow \quad 1 \ \text{ equation} \end{cases}$$

Estimating by OLS the $(T-1)$ equations in orthogonal deviations from individual time-means we obtain the Within Group estimator, i.e. $\hat{\beta}_{wg}$: Estimating by OLS the average equation we obtain the Between Groups estimator, i.e. $\hat{\beta}_{bg}$:

Let

$$\beta_{wg} = E\left(\hat{\beta}_{wg}\right)$$

and

$$\beta_{bg} = E\left(\hat{\beta}_{bg}\right):$$

Rewrite the system as

$$\begin{cases} \overline{y_i} = \overline{x_i}\beta_{bg} + \overline{\varepsilon_i} - \overline{x_i}\beta_{wg} + \overline{x_i}\beta_{wg} \\ y_i^* = x_i^*\beta_{wg} + \varepsilon_i^* \end{cases}:$$

Rearranging, we obtain

$$\left\{ \begin{array}{l} \bar{y}_i = \bar{x}_i \left(\beta_{bg} - \beta_{wg} \right) + \bar{x}_i \beta_{wg} + \bar{\varepsilon}_i \\ y_i^* = x_i^* \beta_{wg} + \varepsilon_i^* \end{array} \right. .$$

Call

$$Y_i^+ = \begin{pmatrix} \bar{y}_i \\ y_i^* \end{pmatrix}; W_i^+ = \begin{pmatrix} \bar{x}_i & \bar{x}_i \\ 0 & x_i^* \end{pmatrix};$$

$$\beta^+ = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_{bg} - \beta_{wg} \\ \beta_{wg} \end{pmatrix}; \varepsilon_i^+ = \begin{pmatrix} \bar{\varepsilon}_i \\ \varepsilon_i^* \end{pmatrix}.$$

The augmented auxiliary model would be

$$Y_i^+ = W_i^+ \beta^+ + \varepsilon_i^+; \qquad i = 1, \ldots, N. \tag{5}$$

If we estimate now β^+ by OLS, we obtain directly the variance of the difference of the two estimators in the upper left part of the variance-covariance matrix of β^+. As first noted by Arellano (1993), under the assumption of spherical disturbances a Wald test on appropriate coefficients in the auxiliary regressions is equivalent to the standard Hausman Test. In order to obtain the HR-test, we perform the Wald test using the White's heteroskedasticity and autocorrelation consistent estimators (White, 1984). This allows us to decide if the difference of the two estimators, namely $\hat{\beta}_{bg}$ and $\hat{\beta}_{wg}$ is significantly different from 0 and the results are robust for the presence of non spherical disturbances. In this context this procedure provides reliable evidence about the presence of correlated effects. It is worthwhile to note that we may lose power in the test but there is often a trade off between strong assumptions and power in testing.

2.3 IVD versus IVL: The HR-Test

If the results of the test at the first stage provide evidence of important measurement errors bias, testing for correlated effects using the comparison between OLS in levels and in deviations, as is the widespread practice, is methodologically not correct and may lead to unreliable results.

An implementation of the HR-test for the presence of correlated effects in presence of measurement errors-in-variables consists in comparing the IVD estimator constructed in the first step of the procedure with the same Generalized Instrumental Variables estimator on the model in levels, hereafter IVL estimator. Under the hypothesis of no correlated effects, an IVL estimator is more efficient than an IVD estimator. However, if correlated effects are present the IVL estimator, which is constructed to control only for a specific structure of the ME (such as for the presence of a period-specific component or for AR(1) ME) loses its consistency while the IVD estimator remains consistent because the transformation of the data used purges the model from the effects of individual-specific components. Therefore, the Hausman framework can be applied. Note that, as in the first step, it is important to construct a number of different tests for correlated effects, one for each possible structure of the measurement errors as indicated by the results of the first step. Once again, the difference between them will be only in the use of different IV estimators. The reasoning of the first step about the choice of valid instruments applies. By analysing the results of a combination of the tests of first and second stage, it is possible to investigate what is the most important source of bias and hence choose the most reliable model specification.

A HR-test for correlated effects in presence of ME can be set out as follows.

As in the comparison between the BG and the WG estimator, we deal with two different estimators that are obtained applying the same estimation method on data transformed in different ways. We obtain the IVL estimator applying the IV methodology to the equations in levels and the IVD estimator applying the IV methodology to the equations in deviations from the mean.

We construct the system

$$\begin{cases} y_i = x_i\beta + \varepsilon_i \\ y_i^* = x_i^*\beta + \varepsilon_i^* \end{cases}$$

Estimating by IV the first group of equations, i.e. the ones in levels, we obtain the IVL estimator, i.e. $\widehat{\beta}_{ivl}$. Estimating by IV the second group, i.e. equations in deviations, we obtain the IVD estimator, i.e. $\widehat{\beta}_{ivd}$.

Let

$$\beta_{ivl} = E\left(\widehat{\beta}_{ivl}\right)$$

and

$$\beta_{ivd} = E\left(\widehat{\beta}_{ivd}\right).$$

Rewrite the system as

$$\begin{cases} y_i = x_i\beta_{ivl} + \varepsilon_i - x_i\beta_{ivd} + x_i\beta_{ivd} \\ y_i^* = x_i^*\beta_{ivd} + \varepsilon_i^* \end{cases}.$$

Rearranging, we obtain

$$\begin{cases} y_i = x_i(\beta_{ivl} - \beta_{ivd}) + x_i\beta_{ivd} + \varepsilon_i \\ y_i^* = x_i^*\beta_{ivd} + \varepsilon_i^* \end{cases}.$$

Call

$$Y_i^+ = \begin{pmatrix} y_i \\ y_i^* \end{pmatrix}; \quad W_i^+ = \begin{pmatrix} x_i & x_i \\ 0 & x_i^* \end{pmatrix};$$

$$\beta^+ = \begin{pmatrix} \beta_{ivl} - \beta_{ivd} \\ \beta_{wg} \end{pmatrix}; \quad \varepsilon_i^+ = \begin{pmatrix} \varepsilon_i \\ \varepsilon_i^* \end{pmatrix}.$$

The augmented auxiliary model would be

$$Y_i^* = W_i^* \beta^* + \mu_i^*; \qquad i = 1, \ldots, N.$$

Estimating the model by IV, we obtain directly the variance of the difference of the two estimators in the upper left part of the covariance matrix of β^*. Unfortunately, in standard econometric packages White's consistent estimators for IV estimators may not be implemented for panel data. In this case, a possible solution can be to obtain the IV estimators as OLS estimators on a further transformed model, as was necessary and explained in the first step of the procedure. After repeating the same steps for the construction of another artificial regression with these new transformed equations and estimating consistently the variance of the OLS estimators, once again a Wald test will allow us to investigate the presence of correlated effects in a reliable way.

3 An Empirical Application: the Use of the Matching Function Framework.

The main purpose of our analysis is to investigate the effects of unobservable factors that, as is well-known, affect two of the variables frequently used in labour market research, namely the stock of unemployment and the stock of vacancies. We use a matching function framework. In recent years the concept of a matching function has been extensively used to explain the working of the labour market.[5] However the majority of the studies are theoretical. Moreover, while the theoretical emphasis is typically on the behavior of microeconomic units, most of the empirical applications have used aggregate data. In recent years, a small number of empirical studies investigating the empirical relevance of the concept at less aggregate levels have been produced. The central question addressed is whether the matching function exhibits constant returns to scale, which is one of the basic assumption in the theoretical literature. In this paper we analyse an empirical matching function but our primary aim is neither an empirical testing of such a stylized relation nor an inspection of the returns to scale exhibited. We will briefly comment on these issues while analyzing the results obtained for the data set under investigation, but the main focus of our empirical analysis is an illustration of the econometric procedure presented in Section 2. The choice of the level of disaggregation and the frequency of the data is guided by the necessity to minimize additional sources of bias. Specifically, we compare different assumptions about unobservable factors that may influence the estimation results and let the data decide what is the most important issue to control for.

[5] See, for instance, Pissarides (2000), Petrongolo and Pissarides (2001) for an extensive review and discussion.

4 Description of the Data and Definition of the Variables.

A longitudinal data set of travel-to-work areas (TTWAs) in the UK observed monthly for the period 1996-2001 has been used. All data are available from the National On-line Manpower Information Service (NOMIS) located at the University of Durham. In the United Kingdom the travel-to-work-areas are considered the standard approximations to self-contained labour markets, i.e. areas in which people both live and work. They are geographic regions with a minimum of 3500 working people where at least 75% of those living (working) in the area should also work (live) there. We consider the most recent TTWAs' definition, based on the journey to work statistics from the 1991 Census of Population. A total of 297 TTWAs are designated in England, Scotland and Wales. Only areas with non missing values[6] are included in the sample used for estimation, reducing the cross section dimension from 297 to 277 areas. Furthermore we also eliminated London, the biggest TTWA, so that we ended up with 276 areas.

The Nomis database contains detailed informations from both sides of the labour market. Unemployment and vacancy data collected by Nomis are registration data provided by local employment agencies (Job Centres). They are administrative data that have the advantage of being readily available on a regular basis, at high frequencies, and at a very disaggregate regional level. Temporal aggregation is an important issue in the estimation of a matching function because it involves estimating flows from stock variables. High-frequency data can mitigate this bias. On the other hand aggregation over space can also be misleading. The estimation of a matching function combining cross sectional and time series observations where the cross section units are the regions may still lead to unreliable results. If the regions have a different size and matching does not take place under constant return to scale, estimates may be affected by a spurious scale effect. Working with TTWAs' data, also this further source of bias should be mitigated, although Burgess and Profit (2001) find evidence for spatial spillovers even at the level of TTWAs.

We use as a proxy for the total number of unemployed the monthly count of claimants who are claiming unemployment benefits on the unemployment count date (second Thursday of each month) and as a proxy for the jobs that are vacant the monthly stock count of notified vacancies that have not been filled at the end of the previous month. The number of vacancies that are filled by job seekers is our measure of total hirings.

We do not arbitrarily adjust the data following, for instance, the correction proposed by Coles and Smith (1996). It is believed that the Job Centre numbers represent approximately one-third of the vacancies and one-quarter of the placings in a TTWA. It is certainly true that registered vacancies are only one channel from which firms recruit personnel but we are not aware of the exact proportions. What if the ratio between measured number of vacancies (or hires)

[6] In twenty TTWAs in the UK there are no data on vacancies available.

and true number of vacancies (or hires) is not constant across areas? We can have a systematic measurement error and completely misleading estimation results. Our approach is to work with the raw data and try to understand what are the most important unobservable factors affecting our data set.

5 Empirical Analysis and Results

We start by considering a standard Cobb-Douglas specification of the matching function in log-linear form:

$$\log M_{it} = \log A + \alpha \log U_{it} + \beta \log V_{it} + \eta_i + u_{it} \qquad \text{Model 1}$$

We indicate by M_{it} the number of hirings in area i during month t; U_{it} and V_{it} the stocks of registered unemployed and of vacancies in area i at the beginning of period t; η_i is a TTWA fixed effect controlling for regional characteristics, including the size of the TTWA; and u_{it} is a white noise random error term. In this framework, α and β are the elasticities of hirings to unemployment and of hirings to vacancies respectively.

Figure 2 contains the graphs plotting different panel data estimates of α and β calculated recursively by adding six months periods. Assuming normality of the estimators, we draw the bands corresponding to a confidence interval of 95%. The hypothesis of constancy is not rejected. If we ignore the odd values of the estimators in the first two years, perhaps affected by administrative changes in the way data had been collected[7], both elasticities appear to be constant in all the models adopted. Therefore the restrictive Cobb-Douglas specification does not seem to be rejected by the data.[8]

Using the diagnostic procedure presented in Section 2, we investigate the effects of different unobservable factors affecting the unemployment and vacancy data by comparing different panel data estimators of the coefficients of the stock of unemployment and vacancies.

Table 1 reports the results for the different panel data estimators involved in the development of the econometric procedure. In our analysis we use IV estimators that control for autocorrelation in the process of the measurement error. We use as instrument for a variable at time t the value of the same variable at time $t - 3$. This is reasonable from a logical point of view because the instrument is the value of the variable at the end of the previous quarter and from a technical prospective because it allows us to control for the presence of measurement errors that follow a moving average process of order one.

[7] This may be related to the government change in 1997.

[8] A transcendental logarithmic model of the matching technology has also been analyzed. The results go beyond the main purpose of this paper. Therefore they are not reported here, but they are available on request.

Table 1: Model 1 -Estimation Results-

Dependent Variable: Log Filled Vacancies

	OLSL	WG	IVL	IVD
Log vacancies¤	0:4295	0:3502	0:5171	0:5425
	(98:48)	(55:47)	(101:27)	(59:11)
Log unempl¤	0:6943	1:4224	0:4450	1:5323
	(4:42)	(1:99)	(2:64)	(1:63)
Const¤	¡ 2:7228	¡ 7:8714	¡ 1:2091	¡ 9:4665
	(¡ 2:28)	(¡ 1:46)	(¡ 0:95)	(¡ 1:34)

¤t di Student in parentheses.

At the …rst stage of the procedure, a HR-test for the equality of WG and IVD gives us evidence of strong measurement errors for both the unemployment data and the vacancy data. The null hypothesis of equality of the two estimators is strongly rejected in both cases ($\hat{A}_1^2$ = 16:40, p = 0:00; $\hat{A}_1^2$ = 15:86, p = 0:00 for unemployment and vacancies respectively). In the second stage of the procedure we investigate the relevance of heterogeneity among areas constant over time. Area speci…c e¤ects on hirings may arise as a result of within-area variation in the matching technology across TTWAs. These technological di¤erences are likely to be correlated with an area size and hence with the area level of unemployment and vacancies. As explained in Section 2, a test for correlated e¤ects in presence of measurement errors in variables consists of comparing IVL and IVD. Applying a HR-test, we cannot reject the hypothesis of equality of the two estimators for the unemployment coe¢cient ($\hat{A}_1^2$ = 1:49, p = 0:2224), but we reject this hypothesis for the vacancies ($\hat{A}_1^2$ = 107:03, p = 0:00). Di¤erent estimation methods controlling for a speci…c kind of bias show di¤erent e¤ects on the coe¢cient of the two variables: our results suggest that the vacancy and unemployment data are contaminated by unobservable factors of di¤erent types. Therefore we can conclude that area speci…c unobservable factors, such as local policies towards the demand or the supply side of the labour market, in‡uence the stock of vacancies but play only a minor role in the determination of the number of unemployed. However, measurement errors remain an important issue to control for.

The lack of a rigorous statistical analysis may lead to a completely di¤erent interpretation of the estimation results. A visual inspection of the table shows that while for the vacancies coe¢cients the discrepancies between di¤erent estimators on the same transformation of the data (OLSL versus IVL and WG versus IVD) are higher than the ones between the same estimators on different transformations of the data (OLSL versus WG and IVL versus IVD), the unemployed coe¢cients show opposite and more marked patterns. There is a huge di¤erence between OLSL and WG and between IVL and IVD. Therefore the more immediate interpretation is to consider the bias due to measurement errors to be the most important problem for the vacancies' coe¢cient, and unobserved heterogeneity bias as the most important one for the unemployment's coefficient. As explained above, this interpretation is not confirmed by the HR-tests. For instance, the particularly marked patterns of the unemployment elasticities may be due to not only the effects of area-specific factors that are neglected in the estimates of the model in levels, but also the presence of strong measurement errors whose effects are magnified in the models in deviations, as is confirmed by the application of the diagnostic procedure.

In order to investigate further the structure of the measurement errors, we estimate the model with time dummies. It is worth noting that the introduction of time dummies is usually used to capture time components of A (efficiency of the matching function) but it allows also for the effects of unobservable factors constant across areas and changing over time. A panel data model which controls for time differences in the technology of matching and one which assumes measurement errors with a time component in an additive structure have the same specification. Either way we investigate whether differences in the intercepts may account for differences in the previous estimators (slopes). The effects of time specific components common to all areas can be relevant in the framework we are considering because it is very likely that the unemployment or vacancies stocks are influenced by nation-wide policies different over time.

We estimate the following model:

$$\log M_{it} = \log A + \alpha \log U_{it} + \beta \log V_{it} + \eta_i + \delta_t + u_{it} \qquad \text{Model 2}$$

where we use the same notation of Model 1. In addition, δ_t is a time specific effect controlling for the influence of temporal factors constant over areas. Table 2 reports the corresponding results from the different panel data methods of estimation presented in Table 1.

Table 2: Model 2 -Estimation Results-

Dependent Variable: Log Filled Vacancies

	OLSL	WG	IVL	IVD
Log vacancies*	0.4299	0.3509	0.5173	0.5434
	(98.51	(55.49)	(101.30)	(59.05)
Log unempl*	0.6635	0.0010	0.4017	1.4881
	(4.14)	(0.00)	(2.35)	(0.49)
Const*	−2.7442	3.0251	−0.9392	−9.7612
	(−1.36)	(0.21)	(−0.45)	(−0.41)

*t di Student in parentheses.

Applying the diagnostic procedure, the hypothesis of equality of WG and IVD is rejected at the first stage for both variables ($\chi^2_1 = 29.84$, $p = 0.00$; $\chi^2_1 = 52.08$, $p = 0.00$ for unemployment and vacancies respectively). This gives us evidence of the presence of measurement errors with a structure different from the one assumed, i.e. additive structure with a time component and with possibly an MA(1) structure, for both variables. At the second stage, we cannot reject the hypothesis of equality of IVL and IVD for the unemployment coe¢cient ($\hat{A}_1^2$ = 1:19, p = 0:3015) but we do reject this hypothesis for the vacancies coe¢cient ($\hat{A}_1^2$ = 104:9, p = 0:00). Therefore the results of the HR-tests are not di¤erent from the ones obtained for Model 1, both in the …rst and in the second stage. Hence time speci…c factors seem to play a minor role. This evidence may suggest that neither side of the labour market, i.e. neither the demand nor the supply, is strongly in‡uenced by national policies.

As in the interpretation of the results in Table 1, conclusions based only on a visual comparison between Tables 1 and 2 may be misleading. An analysis of Tables 1 and 2 shows that while the coe¢cients for the vacancies are almost untouched, there is a striking drop in the WG estimator for the unemployment coe¢cient that cannot be compared to the slight decrease of all the other estimators. The coe¢cient also loses its signi…cance. It seems that, having controlled for area-speci…c and nation-wide time speci…c factors, the e¤ects of the stock of unemployed on the number of hirings are negligible. In other words one could infer that the unemployment data are almost completely explained by these factors. However, this interpretation needs some care. The IVD, robust to measurement errors, does not show such huge bias as the WG but its value is only slightly decreased, as are the estimators for the model in levels. In presence of strong measurement errors in variables, the estimates in the …rst two columns of Tables 1 and 2, namely OLS estimators, are not reliable. They neglect such unobservable factors and may be misleading. Once more it is worth noting that the HR-tests for unobservable heterogeneity in presence of measurement errors have not been applied for the comparison of OLS estimators but IV estimators have been used (third and fourth columns of Tables 1 and 2), as provided by the diagnostic procedure in Section 2.

Therefore a more reasonable interpretation of the results in Tables 1 and 2 suggests that area-speci…c e¤ects and time-speci…c e¤ects, although important for the vacancies data, are not the most relevant unobservable factors. Autocorrelated measurement errors may play a major role.

Nevertheless, we cannot rule out the possibility of other sources of bias arising, for instance, from a measurement error with a multiplicative structure. Indeed, there is evidence that other unobservable factors of di¤erent nature may matter. Using Model 2, OLS estimators on the model in deviations (WG estimators) and IV estimators on the model in deviations (IVD estimators), robust to measurement errors and with a time speci…c component, give di¤erent results for both variables. Therefore, one should be careful in following the widespread practice of choosing a model in deviations if the estimators in levels and in deviations are di¤erent. If the major problem is a measurement issue the most reliable and precise estimators may be OLS in level because the two biases, namely the unobserved heterogeneity bias and the measurement errors bias, have opposite signs and may partially o¤set each other in the model in levels (Griliches and Hausman, 1986). Transforming the model as the WG estimator requires can exacerbate the bias.

Hence, in the matching function framework which we are analyzing the most reliable estimators seem to be the IV estimators that control for measurement errors on the models in levels. The introduction of the time dummies does not make much di¤erence. The estimators of the model in levels present also more reasonable results from a theoretical point of view. The hypothesis of constant returns to scale is not rejected (Model 1: $\hat{A}_1^2 = 0:05$, $p = 0:8211$, Model 2: $\hat{A}_1^2 = 0:23$, $p = 0:6364$).

6 Conclusions

The main implication from these …ndings is a caveat on the empirical use of estimation results in presence of strong unobservable factors in the data set. OLS estimators are almost never reliable but the availability of panel data sets and the use of estimators that control for unobservable heterogeneity bias, as widespread practice, does not always lead to the most reliable results. It is crucial to investigate what is the most important source of bias that a¤ects the data set we are analyzing. Di¤erent kind of unobservable variables may a¤ect data at di¤erent level and dimension of disaggregation. Panel data sets can be helpful in handling these issues. Pooling cross sectional and time series observations, the econometrics of panel data o¤ers a variety of di¤erent estimators for the same parameter, and the behavior of such estimators in the presence of unobserved factors a¤ecting the data may be analyzed. Therefore, it is possible to acquire some knowledge about the kind of errors of speci…cation involved, by checking whether they can actually account for the sign and order of magnitude of the observed discrepancies between estimators. Pursuing such an approach, we propose a methodological development of an econometric procedure aiming to distinguish the e¤ects of di¤erent kinds of unobservable factors on the estimators of the parameters in a panel data model. We have shown that conclusions based only on a visual inspection of the estimation results may be absolutely misleading.

An application of the methodology to investigate widely discussed issues in labour economics is presented. Using a matching function framework, we studied the e¤ects of unobservable factors in the estimated elasticity of hirings to the stock of unemployed and to the stock of vacancies using a panel data set of TTWAs in the UK followed monthly from 1996 to 2001. Other alternatives, such as a di¤erent dependent variable, i.e. unemploymnet out‡ows, or the use of ‡ows variables instead of stock variables have also been investigated but no more satisfactory results obtained.

Our …ndings reveal that the data on unemployment and vacancies are affected by strong systematic measurement errors of arbitrary structure. In this particular case, cross-sectional di¤erences, naturally associated with di¤erent labor market institutions across TTWAs, seem to be important in the determination of the number of vacancies but do not a¤ect strongly the unemployment stocks. However, it is the presence of measurement errors with an unknown structure that plays a major role. Models controlling for unobserved heterogeneity bias may aggravate the measurement error bias. Therefore the most reliable estimators are instrumental variables on the model in levels. The hypothesis of constant returns to scale cannot be rejected.

This investigation does not rule out the possibility that an empirical analysis of the matching function may lead to dissimilar results using a di¤erent data set. For instance, using data disaggregated by age or educational level it is likely that unobservable heterogeneity bias may be the most important issue to control for.

In presence of strong unobservable factors, as it is the case in analyzing the working of the labour market, the choice of the speci…cation of the econometric model to be used is the most important and delicate phase. In our opinion it is often undervalued in empirical studies.

References

Arellano, M. (1993). On the Testing of Correlated E¤ects with Panel Data., Journal of Econometrics, 59, 87-97.

Bowden, R.J. and Turkington, D.A. (1984). Instrumental Variables, Cambridge University Press.

Burgess, S. and Pro…t, S. (2001). Externalities in the Matching of Workers and Firms in Britain, C.E.P. Working Paper, n.± 490.

Coles, M.G. and Smith, E. (1996). Cross Sectional estimation of the Matching Function: Evidence from England and Wales, Economica, 63, 589-597.

Fahr, R. and Sunde, U. (2001). Disaggregate Matching Functions, IZA Discussion Paper, n.± 335.

Griliches, Z. and Hausman, J.A. (1986). Error in Variables in Panel Data, Journal of Econometrics, 31, 93-118.

Hausman, J.A. and Taylor, W.E. (1981). Panel Data and Unobservable Individual E¤ects, Econometrica, 49, 1377-1398.

Hausman, J.A. (1978). Speci…cation Tests in Econometrics, Econometrica, 46, 1251-1271.

Matyas, L., Sevestre, P. (1996). The Econometrics of Panel Data, A Handbook of the Theory and Application, Kluwer Academic Publishers.

Patacchini, E. (2001). On the Use of the Hausman Test with Panel Data, Department of Economics, University of Southampton, Mimeo.

Pissarides, C. (2000). Equilibrium Unemployment Theory, 2nd ed., MIT Press, Cambridge Massachusetts.

Petrongolo, B. and Pissarides, C. (2001). Looking into the Black Box: A Survey of the Matching Function, Journal of Economic Literature, forthcoming.

White, H. (1984). Asymptotic Theory for Econometricians, Academic Press, New York.

Figure 1: Diagnostic Procedure

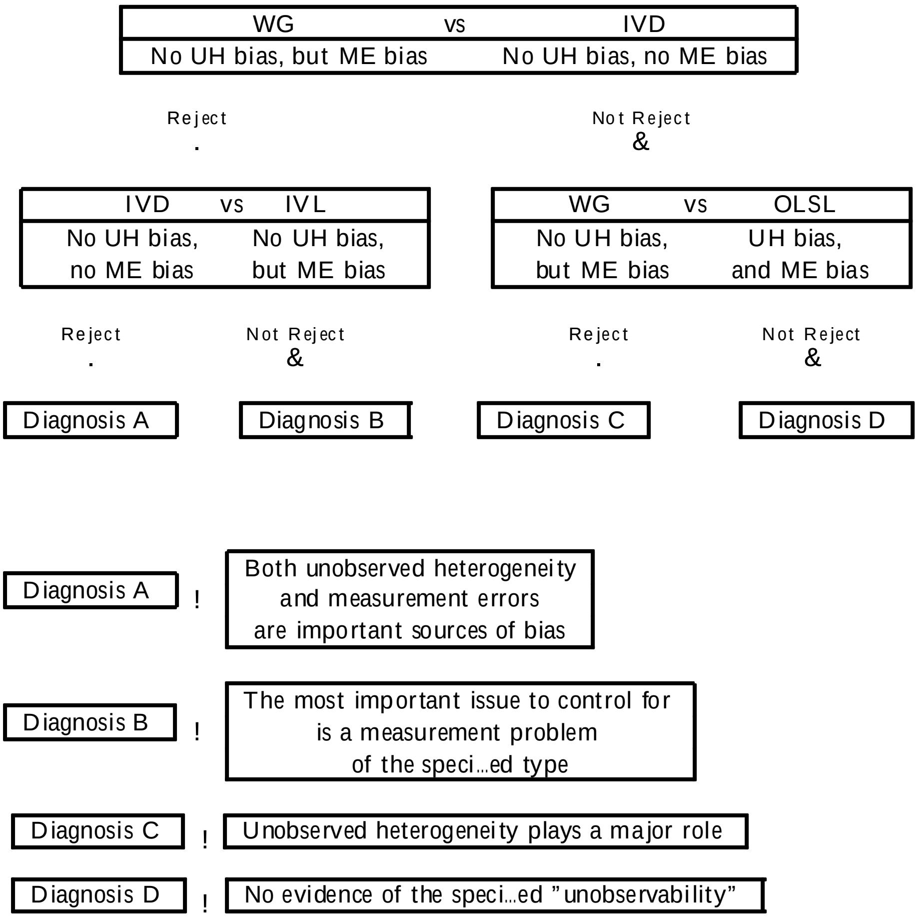
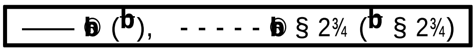
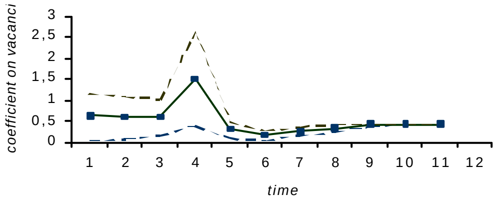
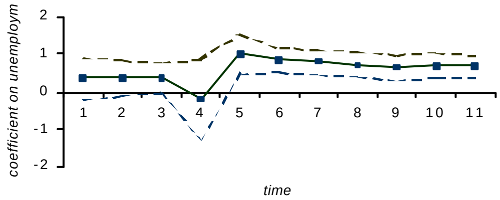


Figure 2: Model 1 -Rolling Elasticities Estimators-

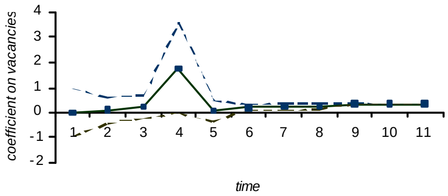




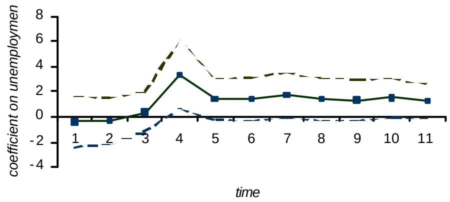
-OLSL estimator-



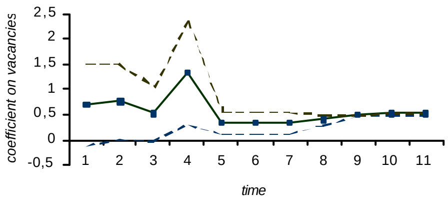
-OLSL estimator-



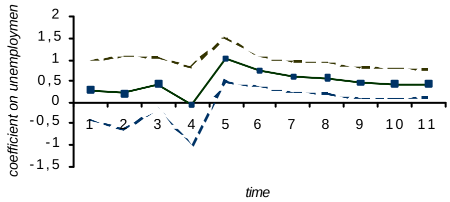
-WG estimator-



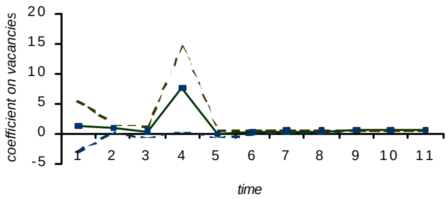
-WG estimator-



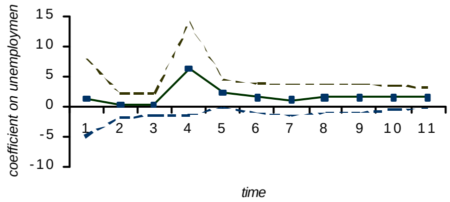
-IVL estimator-



-IVL estimator-



-IVD estimator-



-IVD estimator-