PANEL DATA: MODELS, ESTIMATION, AND THE ROLE OF ATTRITION AND MEASUREMENT ERRORS

Maria Elena Bontempi[1]

[1] Department of Economics
University of Bologna

Many thanks are due to Badi Baltagi, Jacques Mairesse and Jeff Wooldridge for improving my skills, mainly (but not also) during their panel models' lectures at the CIDE Summer Schools. I explicitly acknowledge that their bright ideas inspired this survey.

21th September 2015

Outline of the presentation

- Panel data models: a general specification
- The underlying population model and the sampling scheme that generates data
- The role of heterogeneity and endogeneity
- Estimation methods: how much consistent are the estimates?
- Missing data and measurement errors: some tests

Outline of the presentation

- Panel data models: a general specification
- The underlying population model and the sampling scheme that generates data
- The role of heterogeneity and endogeneity
- Estimation methods: how much consistent are the estimates?
- Missing data and measurement errors: some tests

Outline of the presentation

- Panel data models: a general specification
- The underlying population model and the sampling scheme that generates data
- The role of heterogeneity and endogeneity
- Estimation methods: how much consistent are the estimates?
- Missing data and measurement errors: some tests

Outline of the presentation

- Panel data models: a general specification
- The underlying population model and the sampling scheme that generates data
- The role of heterogeneity and endogeneity
- Estimation methods: how much consistent are the estimates?
- Missing data and measurement errors: some tests

Outline of the presentation

- Panel data models: a general specification
- The underlying population model and the sampling scheme that generates data
- The role of heterogeneity and endogeneity
- Estimation methods: how much consistent are the estimates?
- Missing data and measurement errors: some tests

Panel data models

We can think about a very general model, not only devoted to performance: productivity, financial choices, ownership, investment, failure.... We have an unobserved effects model defined for a large population.

$$y_{it} = \theta\tau_t + \boldsymbol{\beta}' \mathbf{x}_{it} + \boldsymbol{\delta}' \mathbf{q}_i + \underbrace{\mu_i + u_{it}}_{v_{it}} \tag{1}$$

where

- τ_t separate time period intercepts (short T) or time-specific random variables inducing cross-sectional correlation (large T),
- $\mathbf{x}_{it}$ is a $1 \times K$ vector of explanatory variables (maybe also lagged dependent variable) changing with i and t,
- q time-invariant observed variables,
- $\mu_i \sim \mathcal{N}(0, \sigma_\mu^2)$ time-constant unobserved effects randomly drawn along with the observed data,
- $u_{it} \sim \mathcal{N}(0, \sigma_u^2)$ idiosyncratic errors.

This is the usual two-ways error component model: $v_{it} = \mu_i + u_{it}$.

Panel data sample

Usually we assume random sampling in the cross section dimension with large N and fixed $T \Rightarrow$ seeing μ as random is appropriate.

Cross-sectional units are independent, not identically distributed (inid). Since variables are usually correlated over time, we should not assume independence over time, and correlation along time is unrestricted.

This approach also covers cluster samples, where i is replaced by a group index g, while t is replaced by M_g, the number of units in cluster g.
We say that each individual represents a group or cluster: provided that the number of clusters is large relative to the cluster sizes, standard methods can correct for the presence of heteroskedasticity (errors variance not constant over different observations) and/or within-cluster correlation. Fully robust inference is also justified for large cluster size, provided that the number of clusters is also high. See a survey in Wooldridge (2003 AER and 2006 WP).

Panel data models - components

The time period intercepts are parameters that can be estimated. It is convenient, as they:

- take into consideration for correlation over time because of unobservable common characteristics within a firm
- capture common trends in the variation of the dependent variable across cross-sections
- time-demean the data
- the log of nominal values can be used and dummy variables for all time periods (except the base period) capture aggregate price deflators
- remove universal time-related shocks from the errors, thus reducing contemporaneous (cross-sectional) correlation if it is true that all the individuals react in the same manner to macroeconomic events, neighborhood effects, herd behaviour and social norms
- of course, we have collinearity with variables that change over time of the same amount for each unit (age, experience, etc...)

Panel data models - components

The vector $\mathbf{x}_{it}$ can include also interactions of variables with time period dummies, as well as general nonlinear functions and interactions among time-constant and time varying variables, so the model is quite flexible.

The vector $\mathbf{q}_i$ captures unit-specific, time invariant variables: education (years of schooling), type of ownership, belonging to a group, being an exporter, etc...

Panel data models - assumptions

Consistent estimates require that covariates are exogenous, which means uncorrelated with the error term. Here the error term is composed by two parts:

- the unobserved heterogeneity
- the idisyncratic shock.

Denote with $\mathbf{x}_{it}$ all the covariates, and with β the population parameters of interest. Some conditions can defines the partial effect of covariate j on y

$$\beta_j = \frac{\partial E(y \,|\mathbf{x}, \mu)}{\partial x_j}$$

and regard the correlation of the explanatory variables with the two components of the error term.

Panel data models - assumptions and estimating methods

- The POLS is based on contemporaneous exogeneity conditional on the unobserved effect:

$$E(u_{it} | \mathbf{x}_{it}, \mu_i) = 0$$

toghether with uncorrelation between explanatory variables and unobserved heterogeneity:

$$Cov(\mathbf{x}_{it}, \mu_i) = 0$$

Note that not conditioning on μ_i would invalidate the exogeneity assumption.

It rules out standard kinds of endogeneity where some elements of $\mathbf{x}_{it}$ are correlated with u_{it}: measurement errors, simultaneity, time-varying omitted variables.

More importantly, suppose we want to explain productivity with capital inputs and labour force experience (forget at the moment the measurement error...).
Unobserved heterogeneity
- skills of the management/owner, individual motivation, innate ability and intelligence, work ethic, relationships, and innumerable other factors affecting productivity -
is assumed to be constant, unchanged, across the different firms.

For example, the average level of managerial ability is the same regardless of the choice of inputs (or the availability of financial funds, or the effect of belonging to a group, or ...).

Panel data models - assumptions and estimating methods

The (strong) hypothesis is that there are no unobservable characteristics that affect both managerial choices and firms' performance.

$\Rightarrow$ *If this is true*, the slope parameter consistently estimates the economic causality (positive?) going from an additional input to performance.

$\Rightarrow$ *But* it could be that better managers are more able to select good inputs' combinations, i.e. the covariates increase with the average managerial ability.
$\rightarrow$ the covariates are correlated with current shocks v_{it} i.e. they are endogenous with respect to the causal effect β_j.

How to take into consideration for unobserved heterogeneity?

Panel data models - assumptions and estimating methods

FE exploits within deviations i.e. deviations from the "unit-specific" components of the variables which is given by their time averages:

$$y_{it} - y_{i.} = y_{it} - T^{-1} \sum_{t=1}^{T} y_{it}$$

FD exploits first differences of the data:

$$y_{it} - y_{it-1}$$

RE exploits "quasi-time demeaned" data or "lambda" differences, where only a fraction λ of time average is removed:

$$y_{it} - \lambda y_{i.}$$

where

$$\lambda = 1 - \sqrt{\frac{\sigma_u^2)}{T\sigma_\mu^2 + \sigma_u^2}}$$

Panel data models - assumptions and estimating methods

- The FE, FD and RE/GLS, and GMM versions of them, are based on the strict exogeneity conditional on the unobserved effect:

$$E(u_{it} \,|\mathbf{x}_{is}, \mu_i) = 0 \ , \forall t, s = 1, .., T$$

In particular, RE assumes uncorrelation between explanatory variables and unobserved heterogeneity:

$$Cov(\mathbf{x}_{it}, \mu_i) = 0$$

while FE does not pose any restriction on that relationship:

$$Cov(\mathbf{x}_{it}, \mu_i) \neq 0$$

Strict exogeneity rules out lagged dependent variable and feedback effects. So, behaviourally it can often fail: for example, in a production function a firm adjusts her future inputs based on past productivity shocks.

A more reasonable assumption at the basis of dynamic panel data models (usually in FD) is sequential exogeneity or predeterminess conditional on the unobserved effect:

$$E(u_{it} \,|\mathbf{x}_{is}, \mu_i) = 0 \ , s \leq t$$

Since the covariates are correlated with past shocks only, it also allows $\mathbf{x}_{it+1}$ to be correlated with u_{it}. *How to Test*

Panel data models - an unified approach

Based on Mundlak (1978 E) and Chamberlain (1982 JE, 1984 Handbook) idea. "Correlated random effects" (CRE): We model the relationship between μ_i and x_{it}. An attractive assumption is

$$\mu_i = \psi + \boldsymbol{\xi}' \mathbf{x}_{i.} + \nu_i \tag{2}$$

Estimate by RE (or POLS, see Frondel and Vance, 2010 EL):

$$y_{it} = \theta\tau_t + \boldsymbol{\beta}' \mathbf{x}_{it} + \boldsymbol{\delta}' \mathbf{q}_i + \psi + \boldsymbol{\xi}' \mathbf{x}_{i.} + \nu_i + u_{it} \tag{3}$$

This is the variable addition test which unifies FE, RE and BE estimating methods allowing for the robust version of the Hausman test:

- Test H_0: $\boldsymbol{\xi} = 0$ for individual effects uncorrelated with covariates.
- We avoid problems in computing Hausman test
- We can test for sub-sets of covariates
- We can simultaneous estimate the within (effect of an increase over time of **x**) and the between (effect of differences between units of **x**) effects
- The (θ, β) estimates are FE; as a bonus we have estimates of δ.

Panel data models - an unified approach

"Correlated random slopes" (CRS):

$$y_{it} = \theta\tau_t + \beta_i'\mathbf{x}_{it} + \delta'\mathbf{q}_i + \mu_i + u_{it} \tag{4}$$

We can still estimate the population average effect $\beta = E(\beta_i)$

$$y_{it} = \theta\tau_t + \beta'\mathbf{x}_{it} + \delta'\mathbf{q}_i + \mu_i + (\beta_i - \beta)'\mathbf{x}_{it} + u_{it} \tag{5}$$

where $\mathbf{d}_i$=$(\beta_i - \beta)$ is the unit-specific deviation from average.

Panel data models - an unified approach

Suppose that $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}$, with $\mathbf{f}_i$ the unit-specific "level" of the process (that can be approximated by $\mathbf{x}_{i.}$) and $\mathbf{r}_{it}$ the deviations from this level.

- if β_i variation is random across individuals $\Rightarrow$ no bias would emerge, just the model error term will be heteroskedastic and cluster standard errors are recommended
- if β_i variation is related to the average level of $\mathbf{x}_{it}$, i.e. $E(\beta_i \,|\mathbf{x}_{i.}) = E(\beta_i) = \beta$, and strict exogeneity is valid $E(u_{it} \,|\mathbf{x}_{it}, \mu_i, \beta_i) = 0 \Rightarrow$ the β_i variation is absorbed in the fixed effects, the FE estimator is unbiased (but still use cluster standard errors)
- if β_i variation is correlated with idiosyncratic movements in $\mathbf{x}_{it}$, i.e. it is sistematically related with a measure of how far an individual is from the overall/population mean, $(\mathbf{x}_{i.} - \mathbf{x}_{..}) \Rightarrow$ estimates are biased because of correlation of $\mathbf{x}_{it}$ and the error term.

Panel data models - an unified approach

A simple test:

$$y_{it} = \theta\tau_t + \boldsymbol{\beta}_i' \mathbf{x}_{it} + \delta' \mathbf{q}_i + \mu_i + u_{it} \tag{6}$$

$$\mu_i = \gamma' (\mathbf{h}_i - \boldsymbol{\mu}_h) + b_i \tag{7}$$

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \pi' (\mathbf{h}_i - \boldsymbol{\mu}_h) + c_i \tag{8}$$

where $h_i = (\mathbf{x}_{i.}, w_i)$ is a vector of averages of time-varying variables and of time-constant variables that might influence μ_i and β_i Plug in:

$$y_{it} = \theta\tau_t + \boldsymbol{\beta}' \mathbf{x}_{it} + \delta' \mathbf{q}_i + \underbrace{\gamma' (\mathbf{h}_i - \boldsymbol{\mu}_h)}_{\mu_i} + \underbrace{\pi' [(\mathbf{h}_i - \boldsymbol{\mu}_h) \otimes \mathbf{x}_{it}]}_{\boldsymbol{\beta}_i} + b_i + c_i \mathbf{x}_{it} + u_{it} \tag{9}$$

Replace μ_h with sample averages and use OLS/RE or FE to test H_0: $\gamma = 0$ (for CRE) and H_0: $\pi = 0$ (for CRS). Note that:

- if RE is used $c_i \mathbf{x}_{it}$ is removed from the error term (use cluster) but γ is estimated
- if FE is used $\gamma' (\mathbf{h}_i - \boldsymbol{\mu}_h)$ is removed
- β is the average population effect that might be similar to FE without interactions even if interactions are significant
- π measures how the effect of $\mathbf{x}$ on y changes with individual characteristics leading an individual far away from the overall average

Unbalanced panel data

In an unbalanced panel some time periods are missing for some units in the population of interest. Quoting Pudney (1989) panel data exhibit “holes, kinks and corners”. Holes means nonparticipation in some periods; kinks are switching behaviour; corners are nonparticipation at specific points in time. A recent survey on unbalanced panel data is Baltagi and Song (2006 SP).

Sometimes we can have a rotating panel, meaning that randomly some of the original units may be dropped at time $t = k$ and new units added.

Some other times units leave the sample entirely, and usually do not reappear in later years (pure attrition is an absorbing state). More generally, units may reenter the sample after leaving (very complicated case).

A sample selection problem arises if attrition is based on factors that are systematically related to the response variables, even after conditioning on explanatory variables $\rightarrow$ how to test.

Another sample selection problem occurs when people do not disappear from the panel but certain variables are unobserved for at least some time periods (incidental truncation). See Wooldridge (1995 JE), Semykina and Wooldridge (2010 JE).

Unbalanced panel data

Rubin (1976, Biometrika) introduces a distinction:

- Missing at random (MAR): the probability that x_{kit} is missing is independent of its realized value, but may depend on other values, x_{kjt} for $i \neq j$, or on x_{ljt} for $i \neq j$ and $k \neq l$ that we observe.
  $\Rightarrow$ Ignoring missing does not lead to nonresponse bias; observations missing inside clusters of data may be correlated; inefficiency could be tackled by imputation.
- Missing completely at random (MCAR): a special case of MAR, the probability that x_{kit} is missing depends neither on its own values nor on the values of other variables in the data set.
  $\Rightarrow$ Ignoring missing does not lead to nonresponse bias because observed data are a random subsample of the potential full sample. Still inefficiency because data are dropped.

Efficiency loss due to missing data and the seriousness of the problem is directly proportional to the amount of nonresponse (Horowitz and Manski, 1998 JE)

Unbalanced panel data

So, for a random draw i define $\mathbf{s}_i \equiv (s_{i1} \ldots s_{iT})'$ the $T \times 1$ vector of selection indicators $s_{it} = 1$ if $(\mathbf{x}_{it}, y_{it})$ is observed, and zero otherwise. Check if missing observations are on average characterized by systematic features (smallest, services, etc...).
The number of time periods observed for unit i is $T_i = \sum_{t=1}^{T} s_{it}$.
The time-demeaning uses different time periods for different i: $y_{i.} = T_i^{-1} \sum_{r=1}^{T} s_{ir} y_{ir}$.

We can use the "complete case method" i.e. the data for which we observe all of $(\mathbf{x}_{it}, y_{it})$ (or of $(\mathbf{x}_{it}, y_{it}, \mathbf{z}_{it})$ if we need IVs) if selection is strictly exogenous with respect to the idiosyncratic errors, $E(u \,|\mathbf{x}, s) = 0$, with $s = h(\mathbf{x})$ a nonrandom function of exogenous variables $\rightarrow$ ignoring selection will not bias the results.

FE is robust for correlation between s_{it} and the unobserved effects μ_i, while in RE/POLS the μ_i is not eliminated, so we need a stronger assumption about selection being unrelated with the individual effetcs.

Unbalanced panel data

Simple tests in the FE environment.
Add s_{it+1} (and interaction $s_{it+1}\mathbf{x}_{it}$) to the equation at time t. We can also add $(s_{it+1}, s_{it+1}\mathbf{x}_{it+1})$ to check for strict exogeneity of the covariates at the same time.
Or add s_{it-1} (and interaction $s_{it-1}\mathbf{x}_{it}$) to the equation at time t.
Or add a variable equal to the number of periods after period t that unit i is in the sample.
See Nijman and Verbeek (1992, JAE).
It is also a good idea comparing FE (with both within and first differences transformations) estimates on both the unbalanced and balanced panel.

In the RE define dummy variables q_{ir}=1 if $T_i = r$ for $r = 1, .., T-1$ indicating different numbers of time periods, and add them to the model.

Unbalanced panel data

The CRE approach works well in unbalanced panel data when we add any time constant variables to the Mundlak equation (Wooldridge 2010 WP).

The consistency of FE breaks down if the slope coefficients are random but one ignores this in estimation. Since the error term contains $d_i \mathbf{x}_{it}$ where $\mathbf{d}_i = (\beta_i - \beta)$, the selection rule is not exogenous (Note that in the balanced case the FE has some robustness to random slopes).
A simple test of correlation between $\mathbf{d}_i$ and selection through the number of available time periods T_i is to add interaction terms between covariates and dummies for each possible sample size (with $T_i = T$ as the base group) in the equation estimated by FE.

Attrition in panel data

Attrition bias can be tested by including s_{it+1} as an additional explanatory variable, or by using the number of subsequent periods in the sample. First difference is a method that has important advantages for attrition problems, under sequential exogeneity of explanatory variables. Of course, the selection rule in $t+1$ must be unrelated with the error term in t, so selection must satisfy a sort of strict exogeneity assumption and a shock to profit cannot cause a firm to leave the sample. In particular, attrition based on a time-constant unobserved effect can be tackled by first differencing the data. If attrition is not pure, first differencing uses less data than FE.
Forward orthogonalization

$$y_{it} - y_{io} = y_{it} - T_{io}^{-1} \sum_{r=t+1}^{T} s_{ir} y_{ir}$$

where now y_{io} is the average of the observations after t,and $T_{io} = \sum_{r=t+1}^{T} s_{ir}$ is the number of time periods observed after time t,
is an alternative to first differencing and it uses as much data as possible. The use of both first differencing and forward orthogonalization informs about the strict exogeneity of the selection.

Attrition in panel data

Verbeek and Nijman (1996) distinguish between *ignorable* and *nonignorable* selection rules.
In the first case the standard methods can be used to obtain consistent estimates.
In the second case the mechanism that causes the missing observations must be taken into consideration. Vella (1998 JHR) surveys the available methods for estimating models with sample selection bias.

Sometimes, in the hope of mitigating the effects of attrition, panel data sets are augmented by replacing the units that have dropped out with new units randomly sampled from the original population. See Ridder (1992 SCED) and Hirano, Imbens, Ridder and Rubin (2001 E).
When no refreshment is available, the selection rule can be assumed to depend on lagged - but not contemporaneous - variables that have missing data (MAR, Rubin 1976).
Another rule allows the probability of attrition to depend on contemporaneous - but not lagged - variables that have missing data (model of Hausman and Wise 1979 E).

A special issue in spring 1998 of Journal of Human Resources is dedicated to attrition in longitudinal surveys.

Attrition in panel data

An example

The Heckman (1979) or Heckitt procedure estimates:

1. a reduced-form selection equation for $t \geq 2$ by using variables observed at time t for all units with s_{t-1}=1: for example, $\mathbf{x}_{it-1}$, or y_{it-2} in dynamic panel data models, or $\mathbf{x}_{it}$ observed in t when s_{t-1}=1 (age, or other firms' characteristcs). For example,a sequence of probits where in each time period we use the units still in the sample in the previous time period.
2. the equation of interest in which we add the inverse Mills ratios from the $T-1$ cross section probits.

Measurement errors in panel data

Measurement error: despite the explanatory variable x has a well-defined economic meaning, our available measure is wrong in some sense: incorrect response to a survey question; incorrect coding of a correct response; use of a proxy variable for a not observed variable. Some examples in estimating a production function:

- we do not have labour and capital intensity-of-utilisation variables, such as hours of work per employees and hours of operation per machine; we have the number of employees or the (real) value of the capital stock;
- the real values are obtained by using price deflators common across companies (lack of individual prices);
- labour input does not distinguish between blue and white collar; changes in the accounting normative (introduction of IAS) can alter the homogeneity in the definition of capitalized assets
- capital stock could be constructed by a PIM for investment, with some assumptions about the initial value and the depreciation rates.

Measurement errors in panel data

The problem (Solon (1985), Griliches and Hausman (1986)): consider the true model

$$y = bx^c + v$$

However we only observe $x = x^c + \epsilon$ with measurement error $\epsilon \sim \text{iid}(0, \sigma_\epsilon^2)$ and independent of the underlying true value x^c. Hence we estimate

$$y = b(x - \epsilon) + v$$

The model can be rewritten as

$$y = bx + v - b\epsilon = bx + \omega$$

where $Cov(x, \omega) = Cov[(x^c + \epsilon), (v - b\epsilon)] = -b\sigma_\epsilon^2$

$$plim\widehat{\beta} = \beta(1 - \frac{\sigma_\epsilon^2}{\sigma_{x^c}^2})$$

$\widehat{\beta}$ is dowward biased (attenuation bias); the bias does not disappear as $N \to \infty$; the attenuation bias increases as the noise-to-signal ratio $\sigma_\epsilon^2 / \sigma_{x^c}^2$ increases; in case with multiple explanatory variables, the more collinear x^c is with the other explanatory variables, the worse is the attenuation bias. The measurement error also inflates the equation error variance.

Measurement errors in panel data

Comparing within and first difference data transformation gives information about the failure of strict exogeneity because of measurement error.
In within data transformation we have:

$$plim\widehat{\beta}_W = \beta(1 - \frac{T-1}{T}\frac{\sigma^2_\epsilon}{\sigma^2_{x^c_W}})$$

In first difference data transformation we have:

$$plim\widehat{\beta}_{FD} = \beta(1 - \frac{\sigma^2_\epsilon(1-\rho_\epsilon)}{\sigma^2_{x^c}(1-\rho_{x^c}) + \sigma^2_\epsilon(1-\rho_\epsilon)})$$

where $\rho_{x^c} = Corr(x^c_{it}, x^c_{it-1})$ and $\rho_\epsilon = Corr(\epsilon_{it}, \epsilon_{it-1})$.
Biases from random measurement errors (assumed not autocorrelated, $\rho_\epsilon = 0$) are more severe in cases of first difference estimates than in cases of within estimates because first difference magnifies the noise-to-signal ratio independently of T, while within has an inconsistency that shrinks to zero at the rate $1/T$.
The inconsistency becomes very large as $\rho_{x^c} \to 1$; it can be decreased by using "long differences" that are $m > 1$ lags apart ($Corr(x^c_{it}, x^c_{it-m})$ is decreasing in m).

Measurement errors in panel data

A measurement error implies failure to identify the parameter of interest.
The use of proxy variables (instead of omitting the unmeasurable variables) gives smaller bias if the measurement errors are random and independent of the true regression (McCallum, 1972 E).

So, try to include additional information in the model, either in form of additional data or additional plausible assumptions.
Search for additional variables outside the model that can be used as instruments: for example, another measure on x^c, z_{it} with a measurement error orthogonal to the measurement error in x_{is}, all t and s. Under the assumption that the measurement error is not autocorrelated, x_{it-2} and x_{it-3} are valid IVs for Δx_{it}, and so x_{it+1} in a GMM framework (Biorn and Klette, 1998 EL). Biorn (2000 ER) suggests GMM estimators that combine equations in differences and equations in levels; the GMM estimates based on the level equations are more precise than those based on differenced equations. Also see Wansbeek (2001 JE).

Estimation of a production function for Italy

Reference Literature

- Mairesse-Sassenou (1991, NBER, survey), Griliches (1994, AER, US; 1984 and 1998 NBER books on *R&D* and Productivity, various contributions);
- Griliches-Mairesse (1997), Blundell-Bond (2000), Blundell-Bond-Windmeijer (2000, ER) on econometric issues;
- Crepon-Duguet-Mairesse (1998, France), Bond-Harhoff-Van Reenen (2003, Germany and UK), Hall-Mairesse (1995, JE, France), Hall-Mairesse (1996, France and US), Mairesse-Jaumandreu (2005, SJE, France and Spain),Bontempi-Mairesse (2008, NBER, Italy)

The model

$$Q_{it} = A_i B_t L_{it}^{\beta} C_{it}^{\alpha} K_{it}^{\gamma} e_{it}^{\varepsilon} \tag{10}$$

$$(q_{it} - l_{it}) = a_i + b_t + (\mu - 1) l_{it} + \alpha (c_{it} - l_{it}) + \gamma (k_{it} - l_{it}) + \varepsilon_{it} \tag{11}$$

Estimation of a production function for Italy

Reference Literature

- Mairesse-Sassenou (1991, NBER, survey), Griliches (1994, AER, US; 1984 and 1998 NBER books on *R&D* and Productivity, various contributions);
- Griliches-Mairesse (1997), Blundell-Bond (2000), Blundell-Bond-Windmeijer (2000, ER) on econometric issues;
- Crepon-Duguet-Mairesse (1998, France), Bond-Harhoff-Van Reenen (2003, Germany and UK), Hall-Mairesse (1995, JE, France), Hall-Mairesse (1996, France and US), Mairesse-Jaumandreu (2005, SJE, France and Spain),Bontempi-Mairesse (2008, NBER, Italy)

The model

$$Q_{it} = A_i B_t L_{it}^{\beta} C_{it}^{\alpha} K_{it}^{\gamma} e_{it}^{\varepsilon} \tag{10}$$

$$(q_{it} - l_{it}) = a_i + b_t + (\mu - 1) l_{it} + \alpha(c_{it} - l_{it}) + \gamma(k_{it} - l_{it}) + \varepsilon_{it} \tag{11}$$

Table 1: Production function: sample size

Year	$TI = 0$	$TI = 1$	Total	**Year**	$TI = 0$	$TI = 1$	Total
1982	5,146	10,122	15,268	1997	14,075	15,749	29,824
1983	5,101	9,553	14,654	1998	13,786	15,398	29,184
1984	6,371	11,421	17,792	1999	14,251	15,532	29,783
1985	7,286	12,288	19,574	2000	14,394	15,331	29,725
1986	8,084	12,999	21,083	2001	14,138	14,456	28,594
1987	8,490	13,225	21,715	2002	13,276	13,716	26,992
1988	9,044	13,420	22,464	2003	16,469	16,173	32,642
1989	9,922	14,053	23,975	2004	16,875	16,365	33,240
1990	10,563	14,546	25,109	2005	15,929	14,824	30,753
1991	10,421	14,389	24,810	2006	15,088	13,676	28,764
1992	10,328	14,268	24,596	2007	14,115	12,709	26,824
1993	9,275	12,155	21,430	2008	13,226	12,136	25,362
1994	13,216	14,259	27,475	2009	11,958	11,179	23,137
1995	11,198	12,864	24,062	2010	10,529	10,081	20,610
1996	8,111	9,966	18,077	**Total**	**330,665**	**386,853**	**717,518**

Table 2: Production function: statistics

	mean	*p*50	*sd*	*iqr*	*between*	*within*	*residual*	*N*	$\overline{T}$
q_l	3.797	3.791	0.53	0.593	60.34	2.78	36.88	386853	10.13
c_l	3.458	3.488	1.032	1.294	79.54	3.66	16.8	284433	7.54
k_l	0.215	0.246	1.537	1.931	67.17	0.34	32.49	284433	7.54
l	3.908	3.829	1.06	1.242	91.59	0.62	7.8	386853	10.13

Table 3: Production function: pairwise correlations

	q_l	c_l	k_l	l	*inv*	*iinv*
q_l	1					
c_l	0.3612*	1				
k_l	0.1622*	0.0941*	1			
l	-0.0978*	-0.0687*	-0.0626*	1		
inv	0.1428*	0.3281*	0.0479*	-0.0760*	1	
iinv	0.1114*	0.0311*	0.3425*	-0.0316*	0.1117*	1

Table 4: Production function: benchmark and DIFF GMM estimates 1982-1993

				Internal IVs					External IVs				
Var.	OLS	WG	FD	DIF	DIFc	DIFlim	Dpc90	Dpc90T	DIF	DIFc	DIFlim	Dpc90	Dpc90T
c_l	0.153	0.104	0.076	0.053	0.103	0.007	0.138	0.146	0.060	0.067	0.028	0.070	0.092
se	0.003	0.004	0.004	0.031	0.039	0.037	0.037	0.059	0.052	0.049	0.052	0.050	0.052
t	51.0	26.7	18.1	1.7	2.6	0.2	3.8	2.5	1.2	1.4	0.5	1.4	1.8
k_l	0.032	0.004	0.006	0.047	-0.132	0.008	-0.002	-0.009	0.014	0.011	0.010	0.013	0.012
se	0.002	0.002	0.002	0.041	0.098	0.056	0.050	0.065	0.006	0.006	0.006	0.006	0.006
t	18.8	2.4	2.7	1.2	-1.3	0.1	0.0	-0.1	2.2	1.8	1.6	2.1	1.9
l	-0.027	-0.212	-0.548	-0.435	-0.696	-0.640	-0.399	-0.378	-0.682	-0.603	-0.754	-0.667	-0.631
se	0.003	0.008	0.009	0.079	0.178	0.106	0.093	0.115	0.144	0.176	0.168	0.146	0.149
t	-10.8	-26.5	-60.2	-5.5	-3.9	-6.0	-4.3	-3.3	-4.7	-3.4	-4.5	-4.6	-4.2
H	-	-	-	211.7	63.9	113.6	108.0	92.9	115.5	24.3	55.7	70.2	82.6
Hp	-	-	-	0.000	0.000	0.000	0.001	0.001	0.031	0.110	0.008	0.537	0.126
Hdf	-	-	-	142	25	50	66	53	89	17	33	72	69
N	109738	109738	79519	79519	79519	79519	79519	79519	79519	79519	79519	79519	79519
$\overline{T}$	5.07	5.07	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16	4.16

Table 5: Production function: LEV GMM estimates with external IVs

	1982-1993					1995-2010				
Var.	LEV	LEVc	LEVlim	Lpc90	Lpc90T	LEV	LEVc	LEVlim	Lpc90	Lpc90T
c_l	0.207	0.208	0.254	0.210	0.214	0.226	0.244	0.231	0.227	0.214
se	0.016	0.017	0.019	0.016	0.016	0.037	0.064	0.062	0.045	0.041
t	13.2	12.2	13.2	13.2	13.5	6.1	3.8	3.8	5.1	5.2
k_l	0.041	0.040	0.041	0.040	0.038	0.031	0.030	0.032	0.032	0.028
se	0.005	0.005	0.005	0.005	0.005	0.011	0.019	0.018	0.013	0.012
t	8.7	8.9	8.2	8.5	8.3	2.8	1.6	1.8	2.4	2.3
l	0.034	0.036	0.041	0.034	0.039	0.025	0.026	0.019	0.022	0.033
se	0.016	0.014	0.016	0.016	0.016	0.028	0.049	0.046	0.033	0.031
t	2.2	2.5	2.6	2.2	2.5	0.9	0.5	0.4	0.7	1.1
H	118.7	38.4	42.5	106.9	97.8	198.3	32.9	67.2	157.0	155.6
Hp	0.019	0.002	0.124	0.014	0.034	0.180	0.132	0.043	0.154	0.145
Hdf	89	17	33	77	74	181	25	49	140	138
N	109738	109738	109738	109738	109738	156241	156241	156241	156241	156241
$\overline{T}$	5.07	5.07	5.07	5.07	5.07	6.00	6.00	6.00	6.00	6.00

Estimation of a production function for Italy

The example is devoted to show the application of some of the previous discussed issues. In the production function: Q value added; A_i and B_t respectively capture efficiency (unmeasurable firm-specific characteristics, like management ability) and the state of technology (the macroeconomic events that affect all companies, like business cycle and "disembodied technical changes" i.e. changes over time in the rates of productivity growth); labels C, K and L are tangible and intangible capital stocks and labour, respectively, with the associated parameters measuring the output elasticity to each input; ε_{it} is the usual idiosyncratic shocks, allowed to be heteroskedastic and within-firm autocorrelated.

Large and unbalanced panel of Italian manufacturing companies over the period 1982-2010, drawn from the CADS (Company Accounts Data Service of Centrale dei Bilanci, more details in Bontempi and Mairesse [2008]). The total number of observations, more than 717,000, is roughly equally split between services ($TI = 0$) and manufacturing ($TI = 1$) companies; the total number of individuals is 73,072, with the availability of minimum 4 years and of maximum 29 years.

Estimation of a production function for Italy

Focus on manufacturing companies, to produce results in line with those of the literature. In line with the Italian manufacturing division, the data-set is mainly characterized by small and medium-sized firms (with a median number of employees equal to 46 units; about 113 units on average). Input variables are characterized by outliers causing departures of non-parametric measures of spread (inter-quartile range, iqr) from parametric ones (standard deviation, sd). This is particularly evident in intangible capital stock, suggesting that large intangible stocks are concentrated in relatively few companies, and that zeros more prevail here than in the other two inputs. Across companies variability prevails, with shares higher than 60% (in line with the findings in Griliches [1988]).

Temporal span split in two periods, 1982-1993 and 1995-2010, so that we can check the robustness of our findings to changes in the macroeconomic context, as well as to changes in the accounting standards - particularly for the capital stock - following the implementation of the Fourth European Commission Directive since 1993.

Estimation of a production function for Italy

Table 4: some puzzling results (e.g. Mairesse and Sassenou [1991], Griliches [1998]).

- POLS: plausible parameter estimates, in line with factor shares and generally consistent with constant return to scale. But should be biased by omitted heterogeneity and endogeneity, in particular, correlations between covariates and firm-effects (unobservable efficiency levels of companies).
- Controlling for unobserved heterogeneity: less satisfactory parameter estimates.

  > *In empirical practice, the application of panel data methods to micro-data produced rather unsatisfactory results: low and often insignificant capital coefficients and unreasonably low estimates of returns to scale*

  (Griliches and Mairesse [1998] p. 177; also see Mairesse and Hall [1995]).
- FD: affected by random year-by-year noise that hides the signal of data (Griliches and Hausman [1986]); particularly evident in the elasticity of labour (disappointing decreasing returns to scale).
- GMM on FD equations with lagged levels of covariates as IVs; controlling for endogeneity (simultaneous choice of output and inputs) and measurement errors: but overfitting problems, since the number of available IVs depends on the length of the panel (if unbalanced in a complex way) and on the number of endogenous covariates.

Estimation of a production function for Italy

- All the available lags (*DIF*), collapsed (*DIFc*), lag-depth truncation at $t-3$ (*DIFlim*) (Mairesse and Hall [1996] for France and US; Mairesse and Jaumandreu [2005] for France and Spain; Bontempi and Mairesse [2008] for Italy).
- IVs reduction techniques through the principal components analysis (*pca*2, see Bontempi and Mammi, *forth*. SJ): *Dpc*90 (90% of variability explained, separate variables) and *Dpc*90*T* (variables taken together). The principal components extraction produces the best results: overidentifying restrictions not rejected; sensible elasticities of output to capital stocks; from the economic point of view, the reduced form contemplates the possibility of complementarity among productive inputs.
- Usual "internal" IVs: lack of robustness and rejection of overidentifying restrictions;
- "External" IVs: tangible and intangible gross investments (*inv* and *iinv*, see Table 3); preferred over the "internal" ones, for at least one reason: the lags of the explanatory variables may be affected by the same measurement error (possibly correlated over time) that we are trying to tackle.

Estimation of a production function for Italy

- GMM-DIF: the past levels of variables are poor instruments for the current differences, even in a large cross-sectional dimension (see Bound et al. [1995])
- GMM-LEV: under covariance stationarity assumptions of the variable, past differences of investment are used as ("external") IVs for the levels of productive inputs; the relevant information in the variables of interest is kept (see Table 5).
- Estimates are encouraging: robust to changes in the sample periods and in the temporal span; non-rejection by the Hansen test; previous disappointing decreasing returns to scale vanished in favor of constant returns to scale (from an economic point of view, in the first period, or both in economic and statistical terms in the second period); estimated elasticities of output to inputs are consistent with evidence for other countries obtained by using constrained models - like the total factor productivity approach - to avoid endogeneity and GMM estimating problems; good performance of PCA ($Lpc90T$) especially if compared with lag-depth truncation (a reduction strategy commonly adopted in the literature on productivity).

Thank you for your attention!

Panel data models - assumptions and estimating methods

A simple test for lack of strict exogeneity in covariates is:

$$y_{it} = \theta\tau_t + \beta' \mathbf{x}_{it} + \delta' \mathbf{q}_i + \gamma' \mathbf{w}_{it+1} + v_{it} \tag{12}$$

Estimate by FE and test

$$H_0 : \gamma = 0$$

A simple test for lack of contemporaneous exogeneity in covariate $\mathbf{w}$ in:

$$y_{it} = \theta\tau_t + \beta' \mathbf{x}_{it} + \delta' \mathbf{q}_i + \gamma' \mathbf{w}_{it} + v_{it} \tag{13}$$

Estimate by FE the reduce form:

$$\mathbf{w}_{it} = \pi' \mathbf{z}_{it} + \epsilon_{it}$$

Obtain residuals, estimate by FE

$$y_{it} = \theta\tau_t + \beta' \mathbf{x}_{it} + \delta' \mathbf{q}_i + \gamma' \mathbf{w}_{it} + \rho' \hat{\epsilon}_{it} + e_{it} \tag{14}$$

and test

$$H_0 : \rho = 0$$

Go back