Regression models for panel data: Some recent developments

Franco Peracchi

Georgetown University, EIEF and University of Rome Tor Vergata

SOSE Workshop, September 28, 2018

Outline

Introduction

The standard linear model

Extensions of the standard linear model

Failures of exogeneity

Nonlinear models

Big-data and machine-learning methods

Data infrastructures

Why panel data?

Panel (or longitudinal) data consist of repeated observations on a set of units.

My focus is on panels where the units are elements of a well defined population and observations are repeated through time.

This kind of panels combine features of both cross-section and time-series data:

- as for cross-sections, issues of sample design, sample selection and measurement error may affect representativeness of the underlying population;
- as for time-series, the data are naturally ordered by the value of the time index and usually display some regularity or persistence over time.

Advantages of panel data:

- they simplify the analysis of a variety of economic problems that would be much harder to study using a pure cross-section;
- unlike macroeconomic time-series, they allows us to study behavior at the level of the individual microeconomic units, while controlling for time invariant unit-specific unobserved heterogeneity in a flexible manner.

Examples of panel data

Bank of Italy:

- Survey on Household Income and Wealth (SHIW), nationally-representative (almost) biennial survey of households, with a panel component;
- Industrial and Service Firms (“Indagine sulle Imprese Industriali e dei Servizi”), nationally-representative annual panel of firms, stratified by industry and firm size.

SOSE:

- ISA project (replaces the “Studi di Settore” project), nationally-representative annual panel of firms, stratified by industry;
- “Fabbisogni Standard” and “Capacità Fiscale” projects, annual panels of (almost) all units comprising three different levels of government (Municipalities, Provinces and Metropolitan Cities, and Regions).

Issues with panel data

Data collection issues:

- survey design and survey process;
- sample design;
- missing data due to either nonresponse or unbalanced panel design;
- measurement error.

Modeling issues:

- conditions for identifiability of the parameter(s) of interest;
- unobserved heterogeneity;
- nonlinearity.

The standard linear model for balanced panel data

Given T observations $\{(\boldsymbol{X}_{it}, Y_{it})\}$ on an outcome Y and k regressors $\boldsymbol{X}$ for n units (households, firms, municipalities, etc.), the standard linear model for balanced panel data assumes

$$Y_{it} = \alpha_i + \boldsymbol{X}_{it}^\top \boldsymbol{\beta} + U_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \tag{1}$$

where α_i is an unknown unit-specific intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is the vector of parameters of interest, and U_{it} is an unobserved error term.

The "individual effect" α_i represents time-invariant unobserved heterogeneity, i.e. omitted time-invariant determinants of Y.

Different assumptions about the relations between the α_i's, the $\boldsymbol{X}_{it}$'s and the U_{it}'s result in different versions of the standard model.

Conventional linear estimators of $\boldsymbol{\beta}$

Large menu of estimators to choose from:

- Ordinary LS (OLS): Regress Y_{it} on $\boldsymbol{X}_{it}$.
- First differencing (FD): Regress ΔY_{it} on $\Delta \boldsymbol{X}_{it}$.
- Fixed effects (FE) or "within" estimator: Regress $Y_{it} - \overline{Y}_i$ on $\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i$.
- "Between" estimator: Regress $\overline{Y}_i$ on $\overline{\boldsymbol{X}}_i$.
- Generalized least squares (GLS) and feasible GLS (FGLS) estimators: Regress $Y_{it} - \psi \overline{Y}_i$ on $\boldsymbol{X}_{it} - \psi \overline{\boldsymbol{X}}_i$, with ψ known (GLS) or estimated (FGLS) by $\hat{\psi}$.
- Mundlak's correlated RE estimator: Regress Y_{it} on $\boldsymbol{X}_{it}$ and $\overline{\boldsymbol{X}}_i$.
- Chamberlain's correlated RE estimator: Regress Y_{it} on $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}$.

Note: $\Delta \boldsymbol{X}_{it} = \boldsymbol{X}_{it} - \boldsymbol{X}_{i,t-1}$, $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$, $\overline{\boldsymbol{X}}_i = n^{-1} \sum_{i=1}^{T} \boldsymbol{X}_{it}$, and $\overline{Y}_i = n^{-1} \sum_{i=1}^{T} Y_{it}$.

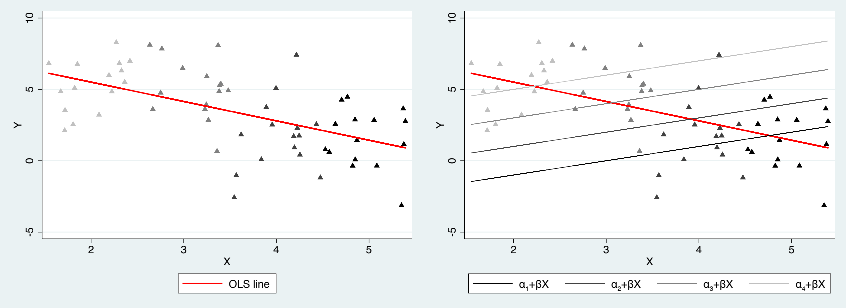
Remarks

The GLS estimator is a matrix weighted average of the FE and “between” estimators.

The GLS and FGLS estimators converge to the OLS estimator when $\mathbb{V}(\alpha_i) \to 0$, and to the FE estimator when $T \to \infty$.

Given a set of instruments, the class of linear estimators of β may be enlarged by considering instrumental variable (IV) versions of all the estimators I mentioned.

Example



10
5
0
-5
Y
2
3
4
5
X
OLS line
10
5
0
-5
Y
2
3
4
5
X
$\alpha_1+\beta X$
$\alpha_2+\beta X$
$\alpha_3+\beta X$
$\alpha_4+\beta X$

Choosing among estimators

Conventional linear estimators are asymptotically normal (Gaussian) under mild regularity conditions, but require different exogeneity assumptions for identification of $\boldsymbol{\beta}$. For example: FD and FE only require mean independence of any U_{it} from $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}$, while "between", GLS and FGLS also require mean independence of α_i from $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}$.

Consequences:

- robustness-efficiency tradeoffs;
- different treatment of time-invariant regressors;
- differences in what one can predict.

With IV procedures an additional issue arises, namely validity (i.e., exogeneity and relevance) of the proposed instruments. In the case of survey data, the characteristics of the interview process and the interviewers provide arguably valid instruments (Nicoletti & Peracchi 2005).

Remarks:

- The Law of Decreasing Credibility (Manski 2003): The credibility of inference decreases with the strength of the assumptions maintained.
- Pre-testing issues arise when using Hausman-type tests of exogeneity assumptions as model selection devices.

Correlated errors

The standard linear model assumes that the errors in (1) are uncorrelated both within and between units. This assumption can easily be weakened to cover cases where the errors are either serially correlated within units or cross-correlated between units.

The second case has become quite relevant as it includes settings that are increasingly common in empirical work:

- clustered samples;
- spatial panel data;
- network panel data.

In all these cases, consistency (or lack thereof) and asymptotic normality of conventional linear estimators of $\boldsymbol{\beta}$ are unaffected.

However, inference is less straightforward because of the more complex nature of the asymptotic variance matrix of the estimators of $\boldsymbol{\beta}$ (Moulton 1986). For this reason, jackknife or bootstrap methods are increasingly used.

Nearly-singular panel designs

The FE estimator, although often preferred because of its weaker identification assumptions, also requires nonsingularity of the second moment matrix

$$\mathbf{S}_{XX} = \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} (\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i)(\boldsymbol{X}_{it} - \overline{\boldsymbol{X}}_i)^\top$$

or, more precisely, requires the smallest eigenvalue of $\mathbf{S}_{XX}$ to be sufficiently far from zero.

Low longitudinal variation of the $\boldsymbol{X}$'s causes failure of this condition and creates both numerical problems and problems for conventional inference, as the usual normal approximation may not be appropriate (Hahn, Ham & Moon 2011)

Possible solution: shrinkage methods to reduce variability of the estimates of $\boldsymbol{\alpha}$ and improve precision of the estimates of $\boldsymbol{\beta}$. Examples:

- quadratic penalty;
- least absolute penalty (Koenker 2004).

Heterogeneous slopes

Why should unobserved heterogeneity be confined to the model intercept?

Replacing $\boldsymbol{\beta}$ in (1) by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$ dramatically increases the number of model parameters from $n + k$ in model to nk.

Mixed models, a parsimonious way of treating heterogeneity in the $\boldsymbol{\beta}_i$, essentially generalize the RE model.

Key assumptions:

- random sample of units;
- mean independence of $\boldsymbol{\beta}_i$ from $\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{iT}$.

Time-varying unobserved heterogeneity

Why should unobserved heterogeneity be time-invariant?

One possibility is to replace α_i in (1) by $\alpha_{i1}, \ldots, \alpha_{iT}$, where the α_{it}'s follow a unit-specific process, typically a stationary or nonstationary ARMA process, e.g. a stationary AR(1) or a random walk. This is equivalent to assuming that the omitted variables in (1) are time-varying instead of time-invariant.

Testing the null hypothesis of time-invariant unobserved heterogeneity in model (1) may be based on the comparison of the FE and FD estimators of $\boldsymbol{\beta}$ (Bartolucci, Belotti & Peracchi 2017).

Further extensions: replace $\boldsymbol{\beta}$ in (1) by $\boldsymbol{\beta}_{it}$.

Time effects and heterogeneous time trends

How about controlling for time-varying "macro" effects?

This just amounts to adding a "time effect" γ_t to (1).

The time effects $\gamma_1, \dots, \gamma_T$ may be modeled in a completely unrestricted way through a sequence of time dummies.

Alternatively, one may assume a low-order polynomial time trend, e.g., a linear one ($\gamma_t = \gamma t$).

Replacing $\alpha_i + \gamma_t$ by a unit-specific linear trend, i.e. by $\alpha_i + \gamma_i t$, is a simple way of modeling heterogeneous time trends. The resulting model is easily estimated via Frisch-Waugh-Lovell, which amounts to linear detrending of both Y_{it} and $\boldsymbol{X}_{it}$.

Failures of exogeneity

The crucial problem when employing model (1) in empirical work is potential failure of the exogeneity assumptions, which may arise from a variety of (not mutually exclusive) reasons:

- omitted variables;
- measurement error;
- simultaneity;
- missing data;
- sample selection.

In what follows I focus on failure of the last two assumptions.

This is not because failure of the first three is less frequent or less relevant, but mainly because failure of the last two leads naturally to nonlinear models, often with an appealing latent linear structure for unobservables plus a nonlinear observation equation mapping unobservables into observables.

Missing data

Common sources of missing data are item and unit nonresponse. Both are widespread and increasingly frequent in sample surveys (Meyer, Mok & Sullivan 2015, Bollinger et al. 2018).

Other sources, specific to panel data and leading to unbalanced panels (i.e. T varying across units), are:

- attrition (monotone or not);
- new entry.

The real issue is not how to allow for unit-specific T's (all conventional estimators allow this), but whether missingness can cause bias.

Classification of missing data mechanisms (Rubin 1976; Little & Rubin 2002):

- missing completely at random (MCAR);
- missing at random (MAR);
- missing not at random (MNAR).

While MCAR and MAR only lead to inefficient estimation of the parameters of interest, MNAR causes bias.

Approaches to MCAR and MAR

Three possibilities:

- Complete-case analysis (not recommended).
- Imputation-based approaches for missing X's:
  - the fill-in approach;
  - missing-indicators methods (Little 1992; Dardanoni, Modica & Peracchi 2011; Dardanoni et al. 2015);
  - multiple imputations (Rubin 1987, 1996).
- Re-weighting approaches for missing Y's:
  - the Horvitz-Thompson method (Horvitz & Thompson 1952);
  - generalizations via inverse probability weighting (Wooldridge 2007).

Approaches to MNAR and sample selection

Two possibilities:

- Achieve point identification of the parameters of interest by explicitly modeling the sample selection process:
  - Heckman framework (Heckman 1979), based on a model of the form (1) for a latent outcome Y_{it}^*, the observability condition $Y_{it} = Y_{it}^*$ if $S_{it} = 1$, and a model for the observability indicator S_{it};
  - more general Tobit models (Amemiya 1984; Vella 1998).
- Do not insist on point identification and only impose "minimal" assumptions that still allow to set identify the parameters of interest, e.g. to assert that $\underline{\mu}_{it} \leq \mathbb{E}[Y_{it}] \leq \overline{\mu}_{it}$:
  - Manski's bounds (Manski 1989) on $\mathbb{E}[Y_{it}]$ for the case when a binary Y_{it} is only observed if $S_{it} = 1$;
  - improving upon Manski's bounds, e.g. by using information on re-entering units when attrition is nonmonotone (Arpino, De Cao & Peracchi 2015).

Linear vs. nonlinear models

Nonlinear models are important when the range of Y_{it} is restricted (e.g. Y_{it} is binary, discrete, categorical, censored or truncated), when data are MNAR, or in the presence of sample selection.

Linear models may still be employed as simple and useful "best approximations" to nonlinear models (Angrist & Pischke 2009).

Key elements of a nonlinear panel data model:

- definition of exogeneity;
- relationship between unobserved heterogeneity and observed regressors;
- temporal dependence among the unobservables.

Difficulties with nonlinear models:

- more complicated identification conditions;
- incidental parameter problem with the FE approach when T is small;
- distinction between model parameters, partial effects at the average, and average derivatives (or average partial effects);
- computational issues;
- need stronger exogeneity assumptions on the distribution of the unobservables.

Nonlinear models with strictly exogeneity

Generalized linear models (McCullagh & Nelder 1989) as a tractable class that includes the Gaussian linear model as special case.

Popular panel data examples include:

- binary logit and probit models (Andersen 1970; Chamberlain 1980);
- ordered logit and probit models;
- Poisson and negative binomial models;
- exponential models.

Approaches to estimation (Liang & Zeger 1986):

- nonlinear LS and nonlinear weighted LS;
- maximum likelihood (ML) and conditional ML (CML).

Nonlinear models with weak exogeneity

The key purpose of models with lagged endogenous variables is to separate true state dependence from spurious state dependence due to unobserved heterogeneity (Heckman 1991).

Role of initial conditions when T is small (Blundell & Bond 1998).

Example: dynamic logit model (Honoré & Kiriazidou 2000; Bartolucci & Nigro 2010, 2012).

Big data

What is “big-data”?

“The leapfrogging of the discourse on big data to more popular outlets implies that a coherent understanding of the concept and its nomenclature is yet to develop” (Gandomi & Haider 2015).

A popular definition is: high-volume, high-velocity, high-variety data (the “Three V” definition). For example: “Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information” (TechAmerica Foundation 2012)

I prefer a narrower definition based on two dimensions of data size: length (n) and width (k). Big data are, at the same time:

- “long”: large n or, in the panel case, large n and T (crucial for validity of asymptotic approximations);
- “wide” (high-dimensional): $k \gg n$, because of either a large number of available variables, or transformations or interactions of the underlying set of variables.

Machine learning methods

This term is often used to denote a rather heterogeneous set of methods employed for analyzing big data.

Their main characteristics include:

- focus on classification or prediction, rather than model fitting and parameter estimation;
- emphasis on flexibility;
- trade-off between bias and variance;
- importance of Bayesian ideas;
- interest in the whole distribution – not just a few features such as mean or variance.

Characterizing a distribution

What characterization of a distribution should one focus on?

The answer partly depends on the nature of Y_{it}. Possibilities include:

- the density function – only defined when Y_{it} is continuously distributed but easily generalized to multivariate outcomes;
- the quantile function – requires no mass points in the distribution of Y_{it} and cannot be generalized to multivariate outcomes;
- the distribution function – can accomodate mass points in the distribution of Y_{it} and is easily generalized to multivariate outcomes.

The last two characterizations are the basis of two alternative methods:

- quantile regression (Koenker & Basset 1978; Koenker 2005, 2017);
- distribution regression (Foresi & Peracchi 1995; Chernozhukhov, Fernandéz-Val & Melly 2013).

Classes of methods

Shrinkage or "regularization" methods (often with a Bayesian justification), including:

- ridge regression (Hoerl & Kennard 1970);
- smoothing splines (Silverman 1985);
- least absolute shrinkage (Frank & Friedman 1993), LAR (Efron et al. 2004) and variants.

Semi-parametric methods, including:

- projection pursuit (Friedman & Tukey 1974; Huber 1985);
- additive modeling (Hastie & Tibshirani 1990);
- neural networks (Ripley 1996) and variants ("deep-learning algorithms");
- local fitting (Tibshirani & Hastie 1987; Fan & Gijbels 1996).

Nonparametric methods, including:

- kernel methods;
- CART (Breiman et al. 1984) and other tree-based methods (bagging, boosting, random forests, etc.).

Inference and model selection

Construction of confidence intervals:

- bootstrap methods (Efron 1980, Efron & Tibshirani 1993): nonparametric and parametric bootstrap;
- subsampling (Politis, Romano & Wolff 1999).

Model selection:

- Mallows C_p (Mallows 1973);
- information criteria: AIC (Akaike 1973), BIC (Schwarz 1978) and variants;
- cross-validation (Stone 1974, 1977);
- the problem of post-selection inference.

Shrinkage and model selection via the LASSO (Tibshirani 1996).

Compromise estimators:

- Bayesian model averaging (Leamer 1978; Raftery, Madigan & Hoeting 1997);
- frequentist model averaging (Hjort & Claeskens 2003; Hansen 2007);
- Bayesian-frequentist fusions, e.g. WALS (Magnus, Prüfer & Powell 2010).

Data infrastructures

Research in the era of big data requires proper infrastructures, such as open data (OD) initiatives and Research Data Centers (RDCs).

Examples of OD initiatives:

- Bank of Italy Remote Access to Micro Data (BIRD) system (https://www.bancaditalia.it/statistiche/basi-dati/bird/).
- VisitINPS Research Program (http://www.inps.it/NuovoportaleINPS/default.aspx?itemdir=47212&lang=IT).
- SOSE Open Civitas portal (https://www.opencivitas.it).

RDCs are partnerships between statistical agencies and leading research institutions. They are secure facilities providing authorized access to restricted-use microdata for statistical purposes only.

An example from the U.S. are the Federal Statistical RDCs, a system of 28 open RDCs locations partnering with over 50 research organizations such as universities (including Georgetown University), non-profit research institutions, and government agencies.

The experience of the Georgetown RDC

The Georgetown RDC, opened in 2017 and located at the Massive Data Institute at Georgetown University's McCourt School of Public Policy, provides secure access to qualified researchers examining a wide range of social and economic issues.

Individuals wishing to conduct research at the Georgetown RDC must submit a research proposal to the Center for Economic Studies (CES) at the U.S. Census Bureau (http://www.census.gov/ces/).

After a researcher has developed a proposal with RDC administrators, the proposal is submitted to the CES for Census review. (Depending on the data sets requested, other agencies might also review the proposal.) Researchers on approved projects must also complete a background investigation. These steps may take months to complete.

See http://mccourt.georgetown.edu/massive-data-institute/RDC?).