

The basic intuition of econometrics: Panel Data, and Qualitative variables

Théophile T. Azomahou
UNU-MERIT, Maastricht University

DEIP, Mar 30 - Apr 3, 2009
Montevideo, Uruguay

Outline of the Lecture

- 1 Set-up
- 2 Linear models for panel data
- 3 Models for qualitative variables (or discrete choice)

Set-up: Panel Data, The Object

- **The Object:** Data that combine time series and cross sections, meaning repeated observations on N individuals over T periods:

$$X_{it} \text{ for } i = 1, \dots, N; \text{ and } t = 1, \dots, T$$

- **Some famous Panel Data sets:** micro and macro panels.
 - ▶ Panel Study of Income Dynamics (PSID, USA)
Collected by the Institute for Social Research, University of Michigan
(<http://psidonline.irs.umich.edu>)
 - ▶ National Longitudinal Survey of Labor Market Experience (NLS, USA)
Sponsored by the Bureau of Labor Statistics
(<http://www.bls.gov/nls/home.htm>)
 - ▶ German Socio-Economic Panel (GSOEP, Germany)
Collected by the German Institute for Economic research, DIW, Berlin
(<http://www.diw.de/soep>)
 - ▶ The Pen World Table, PWT (<http://pwt.econ.upenn.edu>)

Set-up: Panel Data, why should we use it? (a)

Example of issues that could not be studied in either cross-sectional or time-series settings alone.

1. **Labor supply** (Ben-Porath, 1973): At some point in time, in a cohort of women, 50% are working. It is ambiguous whether this implies that, in this cohort, one-half of the women on average will be working or that the same one-half will be working in every period.

These have very different implications for policy and for the interpretation of any statistical results. Cross-sectional data alone are unable to shed any light on the issue.

2. **Production function**: A long-standing problem has been the inability to separate *economies of scale* and *technological change*. Cross-sectional data provide information only about the former, whereas time-series data muddle the two effects, with no prospect of separation.

A panel of data on costs or output for firms can provide estimates of both the rate of technological change (as time progresses) and economies of scale (for a sample of different sized firms at each point in time).

Set-up: Panel Data, why should we use it? (b)

- 3. Program evaluation.** The use of interaction between regressors and time dummies is useful for policy analysis because it allows (partial) effect to change over time.
 - ▶ **Example:** *Natural experiments* on the effect of unexpected property tax reform: 2 groups of individuals
 - Group A: **control group** (houses in a nearby, that are not subject to a property tax reform).
 - Group B: **experimental group** or **treatment group** (houses in the city undergoing unexpected property tax reform), with dummy $dB = 1$ for those in group B and 0 otherwise.
 - Time: consider 2 years data that include the period of the policy change, with dummy $d2$ for the second time (post policy change) period.
 - ▶ **Goal:** analyze the impact of the policy change by interacting $d2$ and dB .

Set-up: Panel Data, why should we use it? (c)

Usefulness of panel data here

Analyze the impact of the policy change by comparing the time changes in the mean of the 2 groups, allowing both for group specific and time specific effects.

- dB (= 1 for those in group B and 0 otherwise) will capture possible differences between the treatment and control groups before the policy change occurs.
- $d2 \cdot dB$ is the interaction term (dummy=1 for those observations in the treatment group in the second year).
- Use what is called **difference-in-difference (DID) estimator**.

However, unbiased **DID estimator** still requires policy change not to be systematically related to other factors that affect the outcome!!!

Set-up: Panel Data, benefits and limitations

Panel Data: benefits

- Controlling for **individual heterogeneity**: panel data suggest that individuals, households, firms, states, countries are *heterogenous*. Risk of bias estimations in pure cross-sections or time-series.
- Panel data provide **more informative data, more variability, less collinearity** among variables, **more efficiency** thanks to large sample size NT , whereas aggregate time-series studies are plagued with multi-collinearity.
- Suitable to study the **dynamics of behavior**: panels can relate the individual's experiences and behavior at one point of time to other experiences and behavior at another point of time.
 - i) the study of the dynamics of transition (Markov process),
 - ii) evaluation of programs needs at least two periods of observation.

Set-up: Panel Data, benefits and limitations

Panel Data: limitations

- *Data collection problem*: coverage issue (incomplete account of population), nonresponse (lack of cooperation of respondent, interviewer error), etc.
- *Measurement errors*: faulty response due to unclear questions, memory errors, deliberate distortion of responses, etc.
- *Selectivity problem*
 - ▶ *Self-selectivity*: if we are only interested in poverty, and people with income larger than 1.5 times the poverty are dropped from the sample (**truncation**, see Hausman and Wise, 1979)
 - ▶ *Attrition*: nonresponse effect is more serious in panel than cross-section, because subsequent waves of the panel are still subject to nonresponse.
- *Short time-series dimension*: in micro panels, asymptotics will rely on N . Increasing time span is costly, and increases chances of attrition.
- *Cross section dependence*: cross-section dependence in macro panels on countries, may lead to misleading inference.

Linear models for panel data

The basic framework for discussion is a regression model of the form

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\alpha} + \varepsilon_{it}$$

with K regressors in \mathbf{X}_{it} , *not including a constant*. The **heterogeneity**, or **individual specific effect** is $\mathbf{Z}_i\boldsymbol{\alpha}$ where \mathbf{Z}_i *includes a constant term*. It can be observed (race, sex, location, etc.), or unobserved (family specific characteristics, individual heterogeneity in skill or preferences, etc.) all of which are taken to be constant over time t .

If \mathbf{Z}_i is observed for all individuals, then the entire model can be treated as an ordinary linear model and fit by least squares.

Remarks: To simplify, i) we will not consider a model with time specific effect, ii) we will assume that the panel is **BALANCED**.

Linear models for panel data

1. **Pooled Regression:** \mathbf{Z}_i contains only a constant term, then OLS provides consistent and efficient estimates of the common α and the slope vector β .
2. **Fixed Effects:** \mathbf{Z}_i is unobserved, but correlated with \mathbf{X}_{it} , then OLS for β is **biased** and **inconsistent**. However, in this instance, the model

$$y_{it} = \mathbf{X}_{it}\beta + \alpha_i + \varepsilon_{it}$$

where $\alpha_i = \mathbf{Z}_i\alpha$, embodies all the observable effects and specifies an estimable conditional mean. This **fixed effects** approach takes α_i to be a group-specific constant term in the regression model.

3. **Random Effects:** If the unobserved individual **heterogeneity**, can be assumed to be uncorrelated with the included variables, then the model may be formulated as

$$y_{it} = \alpha + \mathbf{X}_{it}\beta + u_i + \varepsilon_{it}$$

Estimation of the fixed effects model

The fixed effects specification assumes that differences across units can be captured in differences in the constant term. Each α_i is treated as an unknown parameter to be estimated.

Let \mathbf{y}_i and \mathbf{X}_i be the T observations for the i th unit, \mathbf{i} be a $T \times 1$ column of ones, and let $\boldsymbol{\varepsilon}_i$ be associated $T \times 1$ vector of disturbances.

1. Least Squares Dummy Variable Estimator (LSDV): OLS

Collecting above terms gives

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{i} & \cdots & \mathbf{0} \\ & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{i} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}.$$

$K + n$ parameters to estimate (problem of size and storage capacity when n is large! but no matter)

Estimation of the fixed effects model

2. The WITHIN- and BETWEEN-groups estimators (con'd)

Consider the formulation

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \alpha + \varepsilon_{it}$$

Define the two following transformations:

- the group means,

$$\bar{y}_i = \bar{\mathbf{X}}_i\boldsymbol{\beta} + \alpha + \bar{\varepsilon}_i.$$

- and the deviations from the group means,

$$y_{it} - \bar{y}_i = (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i.$$

where

- ▶ $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$ is the individual mean of y_{it} for individual i
- ▶ $y_{it} - \bar{y}_i$ denotes the deviations from the group means.

Estimation of the fixed effects model

2. The WITHIN- and BETWEEN-groups estimators

- ▶ WITHIN estimator: OLS estimator applied to the deviations from the group means

$$y_{it} - \bar{y}_i = (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i.$$

- ▶ BETWEEN estimator: OLS estimator applied to the group mean transformation

$$\bar{y}_i = \bar{\mathbf{X}}_i\boldsymbol{\beta} + \alpha + \bar{\varepsilon}_i.$$

Estimation of the fixed effects model

3. The First Difference Estimator, FDE

- ▶ The FDE Estimator:

OLS estimator applied to the transformation in first difference,

$$y_{it} - y_{it-1} = (\mathbf{X}_{it} - \mathbf{X}_{it-1})\boldsymbol{\beta} + \varepsilon_{it} - \varepsilon_{it-1}, \quad t = 2, 3, \dots, T$$

- ▶ Interest:

Powerful method for program evaluation. Indeed, applying first difference should difference all variables including binary (dummy) variables indicating participation in a program.

Estimation of the random effects model or error component

Unobserved individual **heterogeneity** are assumed uncorrelated with the regressors. Then the model may be formulated as

$$y_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \alpha + u_i + \varepsilon_{it}$$

where u_i and ε_{it} are random, with:

$$\mathbb{E}[\varepsilon_{it} | \mathbf{X}] = \mathbb{E}[u_i | \mathbf{X}] = 0$$

$$\mathbb{E}[\varepsilon_{it}^2 | \mathbf{X}] = \sigma_\varepsilon^2$$

$$\mathbb{E}[u_i^2 | \mathbf{X}] = \sigma_u^2$$

$$\mathbb{E}[\varepsilon_{it}u_j | \mathbf{X}] = 0 \quad \text{for all } i, t, \text{ and } j$$

$$\mathbb{E}[\varepsilon_{it}\varepsilon_{js} | \mathbf{X}] = 0 \quad \text{if } t \neq s \text{ or } i \neq j$$

$$\mathbb{E}[u_i u_j | \mathbf{X}] = 0 \quad \text{if } i \neq j$$

Application of Generalized Least Squares, GLS !!! (see STATA)

Models for qualitative variables (or discrete choice)

Many settings in which the economic outcome is a 'discrete choice' among a set of alternatives, rather than a 'continuous' measure of some activity. They have in common that they the dependent variable is a dummy of a discrete choice, such as a "yes or no" decision. The general class of models designed for this purpose is **qualitative response** models. And the estimation method is the **maximum likelihood**.

Example of situations

- Number of patents: $y = 0, 1, 2, \dots$. These are **count data**.
- Participation in a program: we equate "no" with 0 and "yes" with 1. These decisions are **qualitative choices**. The 0/1 coding is a mere convenience.
- Decision to innovate (with coding 0/1).
- Opinions of a certain type of legislation: with 0 'strongly opposed,' 1 'opposed,' 2 'neutral,' 3 'support,' and 4 'strongly support.' These numbers are **rankings**. The difference between the outcomes represented by 1 and 0 is not necessarily the same as that between 2 and 1. The **order** matters.

Models for qualitative variables

Three commonly used approaches:

1. The linear probability model, LPM
2. The Logit model
3. The Probit model

The linear probability model, LPM (con'd)

Consider the following simple model:

$$y_i = \alpha + \beta \mathbf{x}_i + u_i, \quad i = 1, \dots, N : \text{firms}$$

where \mathbf{x}_i is the R&D expenditures of firm i , u_i is the error term, and

$$y_i = \begin{cases} 1 & \text{if firm } i \text{ innovates} \\ 0 & \text{if firm } i \text{ doesn't} \end{cases}$$

Models such as the above which express the dichotomous y_i as a linear function of explanatory variables \mathbf{x}_i are called LPM. Why?

Models for qualitative variables

The linear probability model, LPM (con'd)

- Observe that $\mathbb{E}(y_i | \mathbf{x}_i)$: the conditional expectation of y_i given \mathbf{x}_i can be interpreted as the **conditional probability** that the event y_i will occur given \mathbf{x}_i , that is $\mathbb{P}(y_i | \mathbf{x}_i)$.

Indeed, assuming (as usual to obtain unbiased estimator) that $\mathbb{E}(u_i) = 0$ yields

$$\mathbb{E}(y_i | \mathbf{x}_i) = \alpha + \beta \mathbf{x}_i$$

- Denote P_i the probability that $y_i = 1$ (that is, the firm innovates) and $1 - P_i$ the probability that $y_i = 0$ (the firm does not innovate). The distribution of y_i is:

y_i	probability
0	$1 - P_i$
1	P_i

- Then, $\mathbb{E}(y_i) = 0 * (1 - P_i) + 1 * P_i = P_i$. The conditional expectation of the model can be interpreted as the conditional probability:

$$0 \leq \mathbb{E}(y_i | \mathbf{x}_i) = \alpha + \beta \mathbf{x}_i = P_i \leq 1$$

Models for qualitative variables

The linear probability model, LPM: estimation (con'd)

- Shortcomings of the LPM

- ▶ **Normality of disturbances** u_i : not tenable since actually, u_i 's do follow the **binomial** distribution.
- ▶ **Heteroskedastic variances of the disturbances**: Even if $\mathbb{E}(u_i) = 0$ and $\mathbb{E}(u_i u_j) = 0$ for $i \neq j$, it can no longer be maintained that the u_i are homoscedastic. Indeed, we can easily show that

$$\mathbb{V}(u_i) = \mathbb{E}[u_i - \mathbb{E}(u_i)]^2 = \mathbb{E}(u_i^2) = P_i(1 - P_i)$$

with $P_i = \mathbb{E}(y_i | \mathbf{x}_i) = \alpha + \beta \mathbf{x}_i$

- ▶ **Nonfulfillment of $0 \leq \mathbb{E}(y_i | \mathbf{x}_i) \leq 1$** . There is no guarantee that the estimators of $\mathbb{E}(y_i | \mathbf{x}_i)$ in the LPM will necessarily fulfill this restriction (BIG PROBLEM for OLS!).

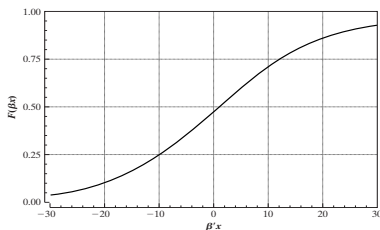
Aim: devise alternatives estimating techniques that guarantee this restriction.

Models for qualitative variables

- Which distribution can do the job?

Any distribution of the form:

- ▶ $\lim_{\mathbf{x}\beta \rightarrow +\infty} \mathbb{P}(y = 1|\mathbf{x}) = 1$
- ▶ $\lim_{\mathbf{x}\beta \rightarrow -\infty} \mathbb{P}(y = 1|\mathbf{x}) = 0$



- Among candidates, the most popular distributions are:
 - ▶ The **Logistic**
 - ▶ The **Probit**

Models for qualitative variables

2. The Logit Model (con'd)

Let us continue with the innovation example. Recall that in explaining the innovation process, the LPM was:

$$P_i = \mathbb{E}(y_i = 1 \mid \mathbf{x}) = \alpha + \beta \mathbf{x}_i$$

where \mathbf{x} is R&D and $y_i = 1$ means that firm i innovates. Now, consider the following representation of innovative firm

$$P_i = \mathbb{E}(y_i = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\alpha + \beta \mathbf{x}_i)}}$$

also to simplify, re-write the (cumulative) **logistic distribution function**,

$$P_i = \frac{1}{1 + e^{-z_i}}$$

where $z_i = \alpha + \beta \mathbf{x}_i$

Models for qualitative variables

2. The Logit Model (end)

- ▶ P_i : probability to innovate
- ▶ $1 - P_i = \frac{e^{-z_i}}{1 + e^{-z_i}}$: probability to not innovate

Then, we have the **ratio**:

$$L = \frac{P_i}{1 - P_i} = \frac{1}{e^{-z_i}} = e^{z_i}$$

which is the **odds ratio** in favor of innovating. For example if $P_i = 0.8$, odds are 4 to 1 in favor innovating firms.

- ▶ Taking the logarithm yields an interesting result:

$$\ln\left(\frac{P_i}{1 - P_i}\right) = z_i = \alpha + \beta\mathbf{x}_i$$

The log of the odds ratio is linear both in \mathbf{x}_i and in parameters.

Models for qualitative variables

3. The Probit Model: latent regression

- ▶ The **Probit** model uses the **normal cumulative distribution function**

$$\mathbb{P}(y = 1|\mathbf{x}) = \int_{-\infty}^{\mathbf{x}\beta} \phi(t) dt = \Phi(\mathbf{x}_i\beta)$$

where $\phi(t) = \frac{1}{\sqrt{\pi}} \exp(-\frac{1}{2}t^2)$ denotes the normal density function.

- ▶ **Innovation example:** assume that the decision of a firm to innovate or not depends on an *unobservable utility index (latent variable)*, y^* , whose sign is known, and which is determined by R&D \mathbf{x}_i as:

$$y_i^* = \mathbf{x}_i\beta + u \quad u_i \sim N(0, \sigma^2)$$

such that the larger the threshold y_i^* , the greater the probability of a firm to innovate:

$$\begin{aligned} y_i &= 1 && \text{if } y_i^* > 0 \\ y_i &= 0 && \text{if } y_i^* \leq 0 \end{aligned}$$

Models for qualitative variables or discrete choice

4. Estimation: The maximum likelihood method (con'd)

Each observation is treated as single draw from a Bernoulli distribution (binomial with one draw).

Let us denote $F(\mathbf{x};\beta)$ a given Cumulative Distribution Function (Logit or Probit). The model with independent observations leads to the joint probability (or likelihood function: the likelihood of having observed the data at hand):

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | \mathbf{X}) = \prod_{y_i=0} [1 - F(\mathbf{x}_i;\beta)] \prod_{y_i=1} F(\mathbf{x}_i;\beta)$$

where \mathbf{X} denotes $[\mathbf{x}_i]_{i=1, \dots, n}$. The likelihood function for a sample of n observations can be conveniently written as

$$L(\beta | \text{data}) = \prod_{i=1}^n [F(\mathbf{x}_i;\beta)]^{y_i} [1 - F(\mathbf{x}_i;\beta)]^{1-y_i}$$

Models for qualitative variables or discrete choice

4. Estimation: The maximum likelihood method (end)

Taking logs, we obtain

$$\ln L = \sum_{i=1}^n \{y_i \ln F(\mathbf{x}_i\boldsymbol{\beta}) + (1 - y_i) \ln[1 - F(\mathbf{x}_i\boldsymbol{\beta})]\}$$

The **likelihood equations** are

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \left[\frac{y_i f_i}{F_i} + (1 - y_i) \frac{-f_i}{(1 - F_i)} \right] \mathbf{x}_i = \mathbf{0}$$

where f_i is the density, $dF_i/d(\mathbf{x}_i\boldsymbol{\beta})$. The choice of a particular form for F_i (**Logit** or **Probit**) leads to the empirical model.

Logit versus Probit: Which is preferable in practice?

Remark: How to chose?

- From a theoretical perspective, the logistic and probit formulations are quite comparable: The logistic has slightly flatter tails. That is the normal curve approaches the axes more quickly than the logistic curve.
- The choice: one of mathematical convenience and ready availability of computer programs. The logistic is generally easier to compute and is generally used.

Thanking you !
Enjoy your STATA