ECON6300/7320/8300
Advanced Microeconometrics
Panel data I

Christiern Rose

[1]University of Queensland

Lecture 5

What are panel data models?

- Models of Panel Data (or longitudinal data) are models for analyzing repeated observations.
- Two dimensional set up is common: data $(y_{it}, \mathbf{x}_{it})$, $t = 1, \ldots T;\ i = 1, \ldots N$, observed at regular time intervals, $T \geq 2$
- Cross-section of time series (hence the **xt** prefix in Stata's panel data commands).
- An additional dimension may be present, e.g. location subscript for spatial dimension.
- Data may come from longitudinal surveys, administrative records of firms/individuals.
- Key feature: repeated observation makes it possible to control for time-invariant unobserved sources of variation.

Uses of panel data models

- Linear panel data models

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

$i = 1, ..., N;\ t = 1, ..., T$ commonly used for outcomes measured on a continuous scale

- Like other econometric models, panel data may be used for:
  - descriptive data modeling
  - causal or structural modeling with endogenous regressors
  - treatment evaluation analysis
- Nonlinear panel data models are applied to discrete ordered or unordered outcomes
- Level of observation is an individual or firm (microeconometrics) or country or industry (macroeconometrics)

Special features of panel data models

- Individual specific heterogeneity is a central theme
  - additive individual specific latent factors
  - non-additive random parameter variation
- Panels are a special case of repeated samples
- Incidental parameter problem and complications of asymptotics
  - large-N-large-T vs large-N-small-T asymptotics
- Panel IV estimation has complicating features
- Nonlinear panel models with correlated effects pose conceptual and computational problems

Panel data set-up

- During observation period each observation unit is uniquely identified
- "Balanced" panels $\Rightarrow$ during the observation period all units are observed, none drop out or are lost
- In microeconometrics the typical set-up involves large N and small (often variable) T ("short panels")
- Initially the data may be in long form ($NT \times (K+1)$) matrix or wide form ($T \times N(K+1)$) matrix
- Observed variation may be due to variation across individuals (gender, ethnicity) or over time (household income), or both
- Panel data analysis simultaneously exploits variation in both dimensions
- Unlike cross section data, out-of-equilibrium behavior can be modelled

Consider data in long form for a two variable regression model

$$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

$$\begin{array}{ccc} y_{11} & 1 & x_{11} \\ y_{12} & 1 & x_{12} \\ .. & .. & .. \\ y_{1T} & 1 & x_{1T} \\ y_{21} & 1 & x_{21} \\ .. & .. & .. \\ y_{N,T-1} & 1 & x_{N,T-1} \\ y_{NT} & 1 & x_{NT} \end{array}$$

Sources of variation in panels

- Two sources of variation – across individuals or across time.
- Total variation of a series x_{it} around its grand mean $\bar{x}$ is

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x})^2 = \sum_{i=1}^{N}\sum_{t=1}^{T}[(x_{it}-\bar{x}_i)+(\bar{x}_i-\bar{x})]^2$$

$$= \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2 + \sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{x}_i-\bar{x})^2,$$

as the cross-product term sums to zero.

- Total sum of squares equals the **within sum of squares** plus the **between sum of squares**.

$$s_W^2 = \frac{1}{NT-N}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2$$

$$s_B^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i-\bar{x})^2.$$

Identification consideration

- A cross-section analysis relies on interindividual differences (variation) in the variable of interest.
  - Example: *How does household spending behavior change as HoH transitions from work to retirement?*
  - Cross-section analysis answers this question by looking at a sample of retirees and non-retirees
  - Panel analysis answers the question by **also** looking at the behavior of households before and after retirement

Identification consideration (2)

- Individuals change their behavior over time in response to changes in their environment
  - Adjustment taking place over time cannot be identified from cross-section data
- Behavior of interest may also change over time as individuals adjust slowly to changes in their environment or because they adapt tochanges

- Individuals may be at different stages of adjustment distributed over time.

- It is possible that individual variation over time contributes substantially to the identification of parameters of interest.

Data example: German Socio-economic Panel (GSOEP)

- A longitudinal panel data set of the population in Germany.
- It is a household based study which started in 1984 and which reinterviews adult household members annually.
- Application to health care: Winkelmann (JHE, 2006); (JAE, 2004)
- In 1996 the cost of healthcare for insured individuals rose as coinsurance rates rose sharply
- Did the number of doctor visits contract significantly (a) immediately, (b) over time?
- Requires data before and after the change from individuals who vary by both observed and unobserved characteristics.

Questions example: Winkelmann (2004)

- Winkelmann (JAE 2004) examines the effect of the healthcare reform enacted in Germany in 1997.
- During 1997, co-payments for prescription drugs increased by 200%, and upper limits were imposed on the reimbursements of physicians.
- What was the effect of the reform on the number of doctor visits? Who was affected and how much?
- A full panel data analysis would confirm whether the reform was effective.
- This data setup is analogous to a *natural experiment*, but issues raised require data from multiple years.

Winkelmann (2004)

Table I. Sample means of doctor visits and selected socio-demographic characteristics, 1995–1999

Year	1995	1996	1997	1998	1999
No. doctor visits	2.687	2.657	2.553	2.353	2.391
(relative change in %)		(−1.1)	(−3.9)	(−7.8)	(+1.6)
No. doctor visits (0/1)	0.348	0.328	0.332	0.372	0.346
Age	38.08	38.20	38.47	38.73	38.92
Unemployed (0/1)	0.085	0.084	0.092	0.085	0.075
Active sport (0/1)	0.295	0.247	0.262	0.307	0.266
Good health (0/1)	0.568	0.562	0.581	0.595	0.580
Bad health (0/1)	0.145	0.138	0.134	0.127	0.129
Observations	6790	6555	6480	6781	6231

Source: German Socio-Economic Panel (GSOEP, $N = 32,837$).

Advantages of Panel Data (1)

- Simplest advantage of panel data $\rightarrow$ increased precision in estimation due to an increase in the number of observations due to combining or **pooling** several time periods of data for each individual.
- **Caveats:**
  - However, individual data may display significant dependence (serial correlation) over time
  - Individuals may not stay within the panel over time which leads to unbalanced panels
  - "Sample attrition" which may lead to selection bias if attrition is not completely random

Advantages of Panel Data (2)

- Panel models allow a better control of unobserved time-invariant heterogeneity than cross section data
  - Microeconometric data display considerable observed and unobserved heterogeneity (UH).
  - **UH may be correlated with observed covariates** $\rightarrow$ need repeated observations per sampled individuals.
- Panel data support learning more about the **dynamics** of individual behavior
  - Can study how individuals react to discrete or continuous shocks
  - Some empirical puzzles require knowledge of parameters that cannot be identified from cross section data
- **Caveats:**
  - Cross-section, within-cluster, and time series dependence reduces information content
  - For robust inference variance calculations must allow for dependence

Dependence concepts

- Temporal dependence for a given i
  - for $\{y_{it}: i = 1, \ldots N; t = 1, \ldots, T\}$ $cov[y_{it} y_{is}] \neq 0$ for $t \neq s$
- Cross section (spatial) dependence
  - $cov[y_{it} y_{jt}] \neq 0$ for $i \neq j$
  - Clustering: $y_{it}^c, y_{jt}^c \subset$ cluster c, $c = 1, \ldots, C$
  - Every observation belongs to just one cluster and there is intra-cluster correlation
  - Cluster can be a household, a district, a class, a school, a village, etc
  - Applies equally to cross section data (i.e. when $T = 1$)
- Important to account for this dependence when estimating the VCE

Limitations of Panel Data

- Getting a population representative continuous panel is not straightforward
- For many possible reasons the initial sample size may be reduced if the participants do not respond to the questions
  - Nonresponse may be total or selective, often involving sensitive issues and privacy concerns ("nonresponse")
  - Panel fatigue after several waves may lead to **nonresponse** ("attrition")
  - Response behavior may be **censored** by death, relocation, lack of access, etc.. ("censoring")
  - Loss through administrative frictions ("frictions")
- Missing observations create gaps and loss of balance

Recurring challenges

1. Model errors likely correlated
2. Parameter identification can depend upon regressor type (exogeneity)
3. Eliminating unobserved heterogeneity may make prediction impossible
4. Some coefficients may vary over time and subpopulations
5. Non-random attrition bias may cause loss of identification of key parameters

Coverage (1)

- Part 1: Basic Linear Panel regression for fixed and random effect models
  - FE vs. RE vs. Pooled models
  - Pooled, RE-GLS, between and within estimators

Coverage (2)

- Part 2: Extensions of the basic model to handle
  - Robust variance estimation
  - Endogeneity and robust IV estimation of FE/RE models
  - Dynamic models, lagged endogenous variables, Arellano-Bond estimator
  - Clustered data and robust variance

Individual effects model (1)

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + \varepsilon_{it}; \quad (1)$$
$$= \mathbf{x}'_{it}\beta + (\varepsilon_{it} + \alpha_i) \quad (2)$$

α_i are random individual-specific effects, and ε_{it} is an idiosyncratic error.

- Two quite different models for the α_i are fixed-effects (FE) and random-effects (RE) models.
- FE vs. RE – potentially misleading terminology because both assume α_i are randomly distributed across i.

- FE : α_i are random across i, but fixed over t for a given i
- RE: α_i are random (exchangeable) across i and t.
- Both assume additively separable α_i;

Individual effects model (2)

- In the FE model α_i in (1) are permitted to be correlated with regressors $\mathbf{x}_{it}$.
  - All time invariant omitted variables
  - Allows a limited form of endogeneity as $E[(\varepsilon_{it} + \alpha_i)|\mathbf{x}_{it}] \neq 0$
    - View the error as $u_{it} = \alpha_i + \varepsilon_{it}$, and permit $\mathbf{x}_{it}$ to be correlated with the time-invariant component of the error (α)
    - But assume that $\mathbf{x}_{it}$ is uncorrelated with the idiosyncratic error component ε_{it}.
- In the RE model the α_i are assumed to be uncorrelated with regressors $\mathbf{x}_{it}$.
  - Two error components $\alpha_i, \varepsilon_{it}$ are uncorrelated with each other
  - α_i induces time-dependence between errors because it is common to T observations $E[(u_{it}u_{is})|\mathbf{x}_{it}] \neq 0$

Incidental parameter problem

- Can fixed effects be estimated consistently?
- One possible estimation method is to jointly estimate $\alpha_1, ..., \alpha_N$ and β.
  - FE model can be written as a model with $N-1$ individual-specific dummy variables
- But for a short panel asymptotic theory relies on $N \to \infty$, and here as $N \to \infty$ so too does the number of fixed effects to estimate.
  - This problem is called the **incidental parameters problem**. Interest lies in estimating β but first we need to control for the nuisance or incidental parameters α_i.
- Solution? - We can consistently estimate β, for time-varying regressors by transforming (1) to eliminate α_i.

Two-way effects model

- A standard extension of the individual effects is a two-way effects model that allows the intercept to vary over individuals and over time:

$$y_{it} = \alpha_i + \gamma_t + \mathbf{x}'_{it}\beta + \varepsilon_{it}. \tag{3}$$

- For short panels it is common to let the time effects γ_t be fixed effects, in which case they are incorporated into $\mathbf{x}_{it}$ by including $T - 1$ indicator variables.

- This model has $N + (T - 1) + \dim[\mathbf{x}]$ parameters that can be consistently estimated if both $N \to \infty$ and $T \to \infty$.

Asymptotics and short panels

- Short panels ($N \to \infty$, T does not).
  - The γ_s can be consistently estimated, so the $(T-1)$ time dummies are simply incorporated into the regressors $\mathbf{x}_{it}$.
  - The challenge then lies in estimating the parameters β controlling for the N individual intercepts α_i.
  - Nickell (1981): bias in short panels from **lagged dependent variables** (i.e. $\mathbf{x}_{it}$ includes y_{it-1})

Pooled Model or Population-Averaged Model

- Pooled models assume that regressors are exogenous and simply write the error as u_{it}, rather than using the decomposition $\alpha_i + \epsilon_{it}$. Then

$$y_{it} = \alpha + \mathbf{x}'_{it}\beta + u_{it}. \quad (4)$$

- Note that $\mathbf{x}_{it}$ here does not include a constant, whereas in cross-section $\mathbf{x}_i$ additionally included a constant term.
- $E[u_{it}\ u_{is}] \neq 0$, for efficient estimation GLS rather than OLS should be used.
- Even if the ϵ_{it} are i.i.d. and uncorrelated with α_i, we have $E[u_{it}\ u_{is}]=E[\alpha_i^2] \quad \forall t \neq s$
- OLS standard errors for PA model are potentially misleading.

Fixed Effects and Random Effects Models

- **Individual-specific effects model** allows each cross-sectional unit to have a different intercept term

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\beta + \varepsilon_{it}, \tag{5}$$

where ε_{it} is iid over i and t. This is a more parsimonious way to express model with any time dummies included in regressors $\mathbf{x}_{it}$

- Throughout make the assumption of **strict exogeneity**

$$E[\varepsilon_{it}|\alpha_i, \mathbf{x}_{i1}, ..., \mathbf{x}_{iT}] = 0, \quad t = 1, ..., T, \tag{6}$$

so that the error term is assumed to have mean zero conditional on past, current and future values of the regressors.

- Strict exogeneity rules out models with lagged dependent or endogenous regressors.

Random Effects (RE) model

- Assumes α_i are random variables that are distributed independently of the regressors.
  - Usually makes the additional assumptions that

$$\begin{aligned} \alpha_i &\sim i.i.d.[\alpha, \sigma_\alpha^2] \\ \epsilon_{it} &\sim i.i.d.[0, \sigma_\epsilon^2] \end{aligned} \tag{7}$$

- More precise terminology: *one-way individual-specific random effect model*, or more simply the *random intercept model*.
- A natural extension is to consider also consider *random slopes* (i.e. $\beta_i \sim i.i.d[\beta, \sigma_\beta^2]$)

RE Equicorrelated Model

- RE model is a specialization of the pooled model, as the α_i can be absorbed into the error term. Then (5) can be viewed as regression of y_{it} on $\mathbf{x}_{it}$ with composite error term $u_{it} = \alpha_i + \varepsilon_{it}$, and (7) implies that

$$\text{Cov}[(\alpha_i + \varepsilon_{it}), (\alpha_i + \varepsilon_{is})] = \begin{cases} \sigma^2_\alpha, & t \neq s. \\ \sigma^2_\alpha + \sigma^2_\varepsilon, & t = s. \end{cases} \tag{8}$$

- RE model imposes the constraint that the composite error u_{it} in (5) is **equicorrelated**, since

$$\rho = \text{Cor}[u_{it}, u_{is}] = \frac{cov[u_{it}, u_{is}]}{\sqrt{var[u_{it}]var[u_{is}]}} = \sigma^2_\alpha / [\sigma^2_\alpha + \sigma^2_\varepsilon] \quad \textit{for} \quad t \neq s$$

  does not vary with the time difference $t - s$.

- Pooled OLS will be consistent but inefficient under the RE model.

GLS estimator of the RE model

- By GMT, the OLS estimator is BLUE if the regression errors are iid and homoskedastic.
- If error variance matrix $\Omega = \mathrm{E}[\mathbf{u}\mathbf{u}'|\mathbf{X}] \neq \sigma^2\mathbf{I}$ and Ω is known and nonsingular, premultiply the linear regression model by $\Omega^{-1/2}$:

$$\Omega^{-1/2}\mathbf{y} = \Omega^{-1/2}\mathbf{X}\beta + \Omega^{-1/2}\mathbf{u}.$$

where $\Omega^{1/2}\Omega^{1/2} = \Omega$. Then

$$\mathrm{V}[\Omega^{-1/2}\mathbf{u}|\mathbf{X}] = \mathrm{E}[(\Omega^{-1/2}\mathbf{u})(\Omega^{-1/2}\mathbf{u})'|\mathbf{X}] = \mathbf{I}_{NT}$$

i.e. errors in this transformed model are zero mean, uncorrelated and homoskedastic.

- So β can be efficiently estimated by OLS regression of $\Omega^{-1/2}\mathbf{y}$ on $\Omega^{-1/2}\mathbf{X}$.

Matrix formulation for GLS when panel data are in long form

$$\mathbf{y}_i = \mathbf{X}_i'\beta + \mathbf{u}_i$$

$$\mathsf{E}[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \Omega = \begin{bmatrix} \mathbf{A}_T & \mathbf{0} & .. & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_T & & \mathbf{0} \\ .. & & ... & .. \\ \mathbf{0} & \mathbf{0} & .. & \mathbf{A}_T \end{bmatrix}$$

$$\mathbf{A}_T = \sigma_\epsilon^2 \mathbf{I}_T + \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T'$$

and $\mathbf{1}_T$ is a T dimensional vector of ones. So Ω is $NT \times NT$.

Between Estimator of the RE model (1)

- Begin with the individual specific effects model. Averaging over all years yields

$$\bar{y}_i = \alpha_i + \bar{x}_i'\beta + \bar{\varepsilon}_i$$

, which can be rewritten as the **between model**

$$\bar{y}_i = \alpha + \bar{x}_i'\beta + (\alpha_i - \alpha + \bar{\varepsilon}_i), \quad i = 1, ..., N, \qquad (9)$$

where $\bar{y}_i = T^{-1}\sum_t y_{it}$, $\bar{\varepsilon}_i = T^{-1}\sum_t \varepsilon_{it}$ and $\bar{x}_i = T^{-1}\sum_t \mathbf{x}_{it}$.

Between Estimator of the RE model (2)

- **Between estimator** is the OLS estimator from regression of $\bar{y}_i$ on an intercept and $\bar{x}_i$.
- It uses variation between different individuals and is the analogue of cross-section regression, which is the special case $T = 1$.

- Between estimator consistent if the regressors $\bar{x}_i$ are uncorrelated with the composite error $(\alpha_i - \alpha + \bar{\varepsilon}_i)$.
  - For consistency we need strict exogeneity, i.e. $E[\varepsilon_{it} | \mathbf{x}_{i1}, ... \mathbf{x}_{iT}] = 0$
  - Rules out FE model since correlation of $\bar{x}_i$ with α_i would lead to an inconsistent estimator

Within Estimator of the RE/FE model

- The within-estimator applies the transformation:

$$y_{it} - \overline{y}_i = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)'\beta + (\epsilon_{it} - \overline{\epsilon}_i)$$

- Estimation of β is by OLS, which is consistent under the FE and RE models.
  - Consistency requires that $(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)$ be uncorrelated with $(\epsilon_{it} - \overline{\epsilon}_i)$, which is true under strict exogeneity (i.e. $E[\epsilon_{it}|\mathbf{x}_{i1}, ..\mathbf{x}_{iT}] = 0$)

- Time-invariant regressors not permitted

First Differences Estimator of the RE/FE model

- The first differences estimator applies the transformation:

$$y_{it} - y_{it-1} = (\mathbf{x}_{it} - \mathbf{x}_{it-1})'\beta + (\epsilon_{it} - \epsilon_{it-1})$$

- Estimation of β is by OLS, which is consistent under the FE and RE models.
  - Consistency requires that $(\mathbf{x}_{it} - \mathbf{x}_{it-1})'$ be uncorrelated with $(\epsilon_{it} - \epsilon_{it-1})$, which is true under strict exogeneity
    - However, it is also a weaker requirement than strict exogeneity, which is particularly useful for dynamic panel data models (next week)

- Time-invariant regressors not permitted

- The variance of the first differences error (for i.i.d. ϵ_{it} it is $2\sigma_\epsilon^2$) is typically larger than that of the within error (for i.i.d. ϵ_{it} it is $\frac{T-1}{T}\sigma_\epsilon^2$), and so we should use within-estimator if we have strict exogeneity.

GLS, Within and Between Estimators

- It can be shown (Maddala, Econometrica 1971) that

$$\begin{aligned}\widehat{\beta}_{GLS} &= [\mathbf{X}'\Omega^{-1}\mathbf{X}]^{-1}[\mathbf{X}'\Omega^{-1}\mathbf{y}] \\ &= [\mathbf{W}_{xx} + \theta\mathbf{B}_{xx}]^{-1}[\mathbf{W}_{xy} + \theta\mathbf{B}_{xy}] \\ \textit{where } \theta &= (1-\rho)/(1-\rho-\rho T) = (\sigma_\varepsilon^2)/(T\sigma_\alpha^2 + \sigma_\varepsilon^2)\end{aligned}$$

- GLS estimator of the RE model is a weighted sum of between and within estimator
- θ measures the weight given to the between variation.

$$\begin{aligned}\theta &\to 0 \Longrightarrow \widehat{\beta} \to \text{Within;} \\ \rho &\to 0 \Longrightarrow \theta \to 1 \Longrightarrow \widehat{\beta} \to \text{Pooled OLS}\end{aligned}$$

- ρ is large only when between variation is large

Limitations

- FE model
  - $E[y_{it}|\alpha_i, \mathbf{x}_{it}] = \alpha_i + \mathbf{x}'_{it}\beta$, assuming $E[\varepsilon_{it}|\alpha_i, \mathbf{x}_{it}] = 0$, so $\beta_j = \partial E[y_{it}|\alpha_i, \mathbf{x}_{it}]/\partial x_{j,t}$.
  - The FE model obtains a consistent estimate of the marginal effect of the j^{th} regressor on $E[y_{it}|\alpha_i, \mathbf{x}_{it}]$ (provided $x_{j,it}$ is time varying) even if regressors are endogenous (i.e. $cov(\mathbf{x}_{it}, \alpha_i) \neq 0$).
  - Knowledge of β does not give complete information on the process generating y_{it}. In particular for prediction we need an estimate of $E[y_{it}|\mathbf{x}_{it}] = E[\alpha_i|\mathbf{x}_{it}] + \mathbf{x}'_{it}\beta$, and $E[\alpha_i|\mathbf{x}_{it}]$ cannot be consistently estimated in short panels.
- RE model
  - Assumed that α_i is purely random, a stronger assumption that implies that α_i is uncorrelated with the regressors (i.e. $cov(\mathbf{x}_{it}, \alpha_i) = 0$).