Regression models and panel data
Part I: Introduction

Rodolfo Metulini

✉ rmetulini@unisa.it

Department of Economics and Statistics (DISES) - University of Salerno

Outline

Introduction

- student's presentation
- presentation of myself
- presentation of the short course
- syllabus, textbook and material
- presentation of the examination method

Panel data

Definition

Panel data: the pooling of observations on a cross-section of households, countries, firms, etc. over several time periods. This can be achieved by surveying a number of firms, households or individuals and following them over time (Baltagi, 2008)

Big data

Big data are characterized for their:

1. Volume,
2. velocity,
3. variety,
4. veracity and
5. value.

Panel data & Big data

- Panel data mainly addresses the issue of data heterogeneity along space and time
- Heterogeneity is connected to many V's because:
  1. heterogeneity emerges with high volume of data,
  2. veracity (uncertainly) increases when both space and time index are considered,
  3. to produce the panel dataset it may be needed to extract data from many sources (variety),
  4. the value is increased, as heterogeneity means that more informations can be extracted from data

What we do and we do not do in this course

We investigate the (causal) relation among variables in a panel assuming:

- that the true relation between the variables is **linear**
- that the dependent variable (Y) is **quantitative**, so that we can adopt **normal** distributions

We do not study models adopted when:

- we have to assume a non linear relation between variables
- the dependent variables is a count data, so that we have to model it by a Poisson, or a Binomial distribution

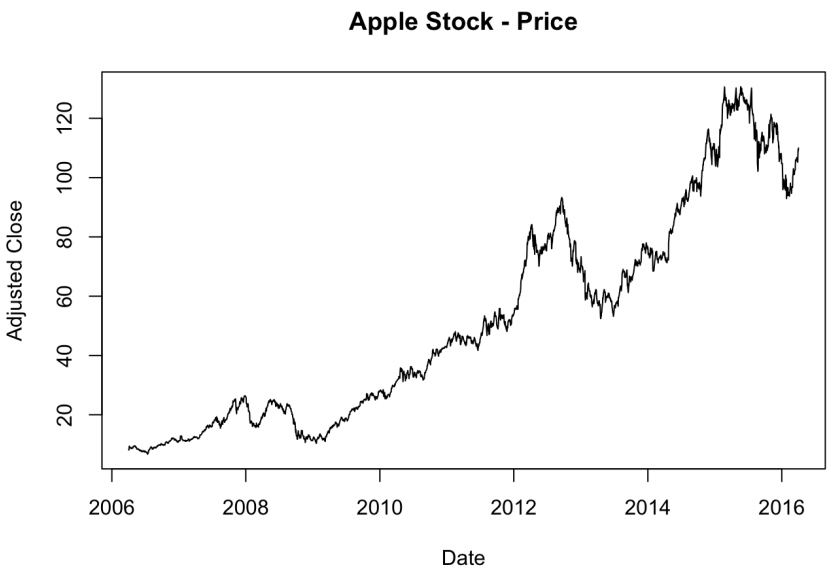
Many books addressing models for non linear relations and/or count data exists, but here we limit our attention to linear relations among (possibly) normal variables

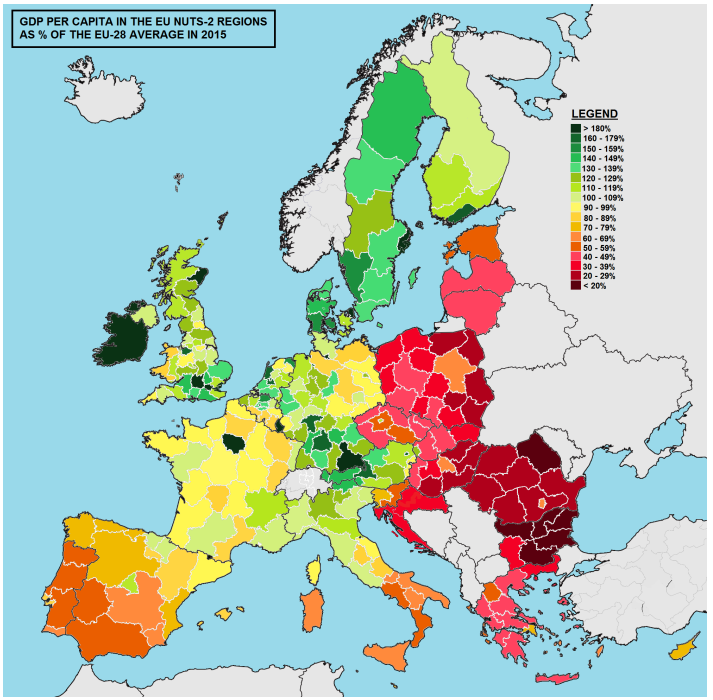
Cross sections and time series (i)

- In statistics and econometrics, **cross-sectional** dataset is a collection of one or more variables for a sample of the population which observes different individuals in just one (and the same for all individuals) period of time, disregarding time.
- In statistics and econometrics, a **time-series** dataset is a collection of one or more variables for one individual along a collection of ordered periods of time
- with time series we can highlight variation in time, with cross section variation inter-individuals
- Examples of time series: i) real GDP by trimester, from Q1.2007 to Q.4.2020. ii) Daily variation of Unicredit on Stock exchange market.
- Example of cross sections: i) the average labour cost per capita per hour for a sample of Chinese firms producing rice. ii) the GDP growth for NUTS2 regions in the period 01-01-2020 to 31-12-2020.

Cross sections and time series (ii)

time series highlight variation in time, cross sections display inter-individuals variations (or heterogeneity)





Panel data

Definition

Panel data: the pooling of observations on a cross-section of households, countries, firms, etc. over several time periods. This can be achieved by surveying a number of firms, households or individuals and following them over time (Baltagi, 2008)

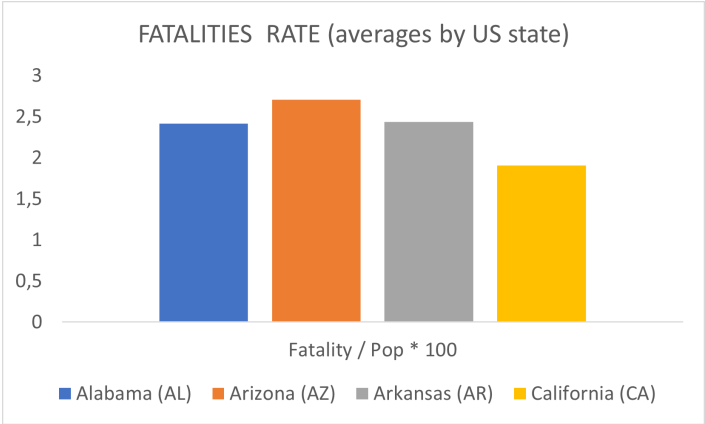
- Panel data allow to account for both time and individual variability
- May regards macro or micro phenomena, so, individuals may be firms or regions
- Panel may be balanced (same individuals along different periods) or unbalanced (the sample of individuals changes along time)

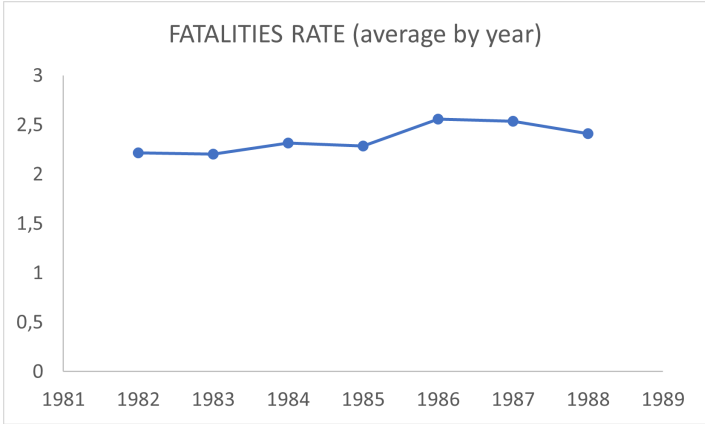
Panel data structure: an example

	state	year	beertax	frate
1	al	1982	1.53937948	2.12836
2	al	1983	1.78899074	2.34848
3	al	1984	1.71428561	2.33643
4	al	1985	1.65254235	2.19348
5	al	1986	1.60990703	2.66914
6	al	1987	1.55999994	2.71859
7	al	1988	1.50144362	2.49391
8	az	1982	0.21479714	2.49914
9	az	1983	0.20642203	2.26738
10	az	1984	0.29670331	2.82878
11	az	1985	0.38135594	2.80201
12	az	1986	0.37151703	3.07106
13	az	1987	0.36000001	2.76728
14	az	1988	0.34648702	2.70565
15	ar	1982	0.65035802	2.38405
16	ar	1983	0.67545873	2.39570
17	ar	1984	0.59890109	2.23785
18	ar	1985	0.57733053	2.26367
19	ar	1986	0.56243551	2.54323
20	ar	1987	0.54500002	2.67588
21	ar	1988	0.52454287	2.54697
22	ca	1982	0.10739857	1.86194
23	ca	1983	0.10321102	1.80672
24	ca	1984	0.09890110	1.94611
25	ca	1985	0.09533899	1.88128
26	ca	1986	0.09287926	1.94548
27	ca	1987	0.09000000	1.98966
28	ca	1988	0.08662175	1.90365

Panel data structure: time and individual heterogeneity

- With panel we have more than 1 information per year and more than 1 information per state
- Computing averages by state and by year we can highlight, respectively, the presence of inter-state heterogeneity and inter-year heterogeneity





- Each state presents a different level
- each year presents a different level as well.

Role of individual heterogeneity

- The main interest when using regression models is that of correctly studying the causal relation between Y (the dependent) and X (the independent). In doing so, it is important to correctly account for individual heterogeneity in Y.
- Stock and Watson (2012) offered an example:
- The research question is whether taxing alcoholics can reduce deaths due to road's incidents
- $frate_n = \alpha + \beta beertax_n + e_n$ estimated on year 1982 returns a **positive** β (!!!). $frate_{nt} = \alpha + \beta beertax_{nt} + e_{nt}$ estimated on the full panel, returns a positive β as well!
- the model $frate_{nt} = \alpha_n + \beta beertax_{nt} + e_{nt}$ accounts for individual heterogeneity via α_n (state-level parameter). This model returns a **negative** β

Role of individual heterogeneity (discussion)

- WHY?? unobserved state-level characteristics (not included in the first two models) are correlated to local beer tax.
- If I miss to include something relevant in the model, and this is correlated with X, OLS is biased and inconsistent.
- Can I include state-level characteristics in cross section model?
- NO, because it involves the estimation of n parameters (related to the n intercepts) on n observations (no degrees of freedom)
- I might know a measure for state-level characteristics, but they are generally unknown
- TAKE HOME MESSAGE: By using cross section data, one may obtain erroneous results (about the causal effect of a regressor on the dependent).

Application

Application in R: example 1.1 (Croissant, Millo) (solution 2 only)

Benefits from using panel data

1. Controlling for individual heterogeneity;
2. more informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency;
3. better able to study the dynamics of adjustment;
4. better able to identify and measure effects that are simply not detectable in pure cross-section or pure time-series data;
5. allow to construct and test more complicated behavioral models than purely cross-section or time-series data.

Controlling for individual heterogeneity

- Panel data suggests that individuals, firms, states or countries are heterogeneous.
- Time-series and cross-section studies not controlling this heterogeneity run the risk of obtaining biased and inconsistent results.
- Baltagi and Levin (1992) consider cigarette demand across 46 American states for the years 1963–88 (t=26):
  $cons_{nt} = cons_{n,t-1} + price_{nt} + income_{nt} + e_{nt}$.
- The model does not consider unobservable time invariant Z_n (e.g., religion, education) or state invariant W_t (e.g., advertising on TV).
- Authors shaw that, omitting Z_n and/or W_t, results may be biased.
- Panel data is able to control for these unobserved variable by including individual- and time-specific effects

More informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency

- Time-series are plagued with multicollinearity; for example, in the case of demand for cigarettes there is high collinearity (reminds linear regression assumptions) between cigarettes' price and income (considering US aggregated data)
- In panel data collinearity is less likely, because the variation in data can be decomposed in within-states and between-states (usually bigger than within)
- With panel data, having larger samples, it is possible to estimate more complex models (with more parameters)
- For example, it may be possible to estimate a state-varying parameters model $y_{nt} = \alpha + \beta_n x_{nt} + e_{nt}$

better able to study the dynamics of adjustment

- Unemployment, job turnover, poverty, growth, etc.. (which presents a cyclic trend with a cycle's duration) are better studied with panels.
- If these panels are long enough, they can shed light on the speed of adjustments to economic policy changes (e.g, elasticity of the price of a cup of coffee on price of inputs after an increase in taxation).
- For example, differently from cross sections and time series, panel data:
  1. can estimate what proportion of those who are unemployed in one period can remain unemployed in another period;
  2. enables to determine at what extent countries' employment rate in time t is benefiting from a government policy in $t-1$;
  3. allow to determine which pharmaceutical firms are benefiting from an increase on EU research funds.

better able to identify and measure effects that are simply not detectable in pure cross-section or pure time-series data

- Example: suppose that we have a cross-section of women with a 50% average yearly labour force participation rate.
- This might be due to:
  1. each woman having a 50% chance of being in the labour force, in any given year
  2. 50% of the women working all the time and 50% not at all.
- Case 1 has high working turnover, while case 2 has no working turnover: only panel data could discriminate between these cases

allow to construct and test more complicated behavioral models than purely cross-section or time-series data

- With cross sections we are forced to treat individuals as all having the same behaviour (e.g., all firms' productivity react to EU research funds with the same elasticity)
- With panel data we can treat individuals as having different behaviours, since we are allow to model a firm-varying coefficient model (so that, funds' elasticity is different along firms)
- Moreover, firm's productivity at time t may depends on the productivity in time $t-1$ (dynamic models),
- or it may depends on the productivity of the neighbours (firms located close by) (spatial models)

Limitations from using panel data

Better to say *problems with data collection*

- Design and data collection problems;
- distortions of measurement errors;
- Selectivity problems:
  1. self-selectivity;
  2. nonresponse;
  3. attrition.
- short time-series dimension;
- cross-section dependence.

Design and data collection problems

- These issues include:
  1. problems of coverage (incomplete account of the population of interest)
  2. nonresponse (due to lack of cooperation of the respondent or because of interviewer errors)
  3. recall (respondent not remembering correctly)
  4. frequency of interviewing and interview spacing (reference period)
- For an extensive discussion of problems that arise in designing panel surveys as well as data collection and data management issues see Kalton et al. (1989)

Distortions of measurement errors

- Measurement errors may arise because of:
  1. faulty responses due to unclear questions
  2. memory errors
  3. deliberate distortion of responses (e.g., prestige bias)
  4. misrecording of responses
  5. interviewer effects
- **Panel advantage**: Cross-section data users have little choice but to believe the reported information in the survey (unless they have external information) while users of panel data can check for inconsistencies of responses along different interviews.

Selectivity problems

- **Self selectivity**: Example on people's wage: some people choose not to work because the reservation wage is higher than the offered wage.

  We will observe the characteristics of these individuals but not their wage: the sample is going to be truncated (when data is missing) or censored (when we just know wage is under a threshold)

- **Nonresponse**: refusal to participate, nobody at home, untraced sample unit, etc...

  Partial nonresponse occurs when one or more questions are left unanswered.

  Complete nonresponse occurs when no information is available from the sampled individual

- **Attrition**: Nonresponse is a more pronounced issue in panel (compared to cross section).

  Subsequent waves of the survey are still subject to nonresponse because respondents may die, or move, or find that the cost of responding is high.

Short time series dimensions

- Typical micro panels involve data covering a short time span for each individual.
- This means that asymptotical consistency relies on the number of individuals tending to infinity.

Cross sectional dependence

- Macro panels on countries or regions with long time series that do not account for cross-country dependence (the variable Y is not *iid*) may lead to misleading inference.
- Accounting for cross-section dependence turns out to be important and affects inference.
- Panel unit root tests are suggested that account for this dependence (Baltagi - 2008, ch. 12)

Unobserved heterogeneity

- From a statistical point of view, panel data addresses the issue of **unobserved heterogeneity** (i.e., controlling for it to avoid biased estimations)
- Let consider the model $y = \alpha + \beta x + \gamma z + e$, where x is an observable regressor and z is unobservable.
- The "feasible" model is $y = \alpha + \beta x + e$ and may suffer for omitted variable bias: OLS $\hat{\beta}$ is consistent and unbiased if z is uncorrelated with either x or y
- Example: agricultural production function (Mundlak, 1961). output (y) depends on labour (x) and soil quality (z). Soil quality is correlated with the effort (labour), hence $\hat{\beta}_{OLS}$ will be inconsistent for β.

Panel data solution

- The panel data model aims to avoid this problem

$$y_{nt} = \alpha + \beta^T x_{nt} + (\mu_n + \upsilon_{nt}) \tag{1}$$

  where μ_n is time-invariant (and it represents the unobservable characteristics on the individuals)

- the objective is to eliminate (wipes out) μ_n from the model.

Eliminating the unobservables

- If unobservable characteristics are time invariant, μ_n is a good proxy for it, and it is possible to rewrite the model in eq. 1 in terms of observables only.

  1. Differencing method (then, OLS method, which is consistent, since we **remove Z**):

  $$\Delta y_{nt} = \beta^T \Delta x_{nt} + \Delta \mu_n + \Delta v_{nt} \quad (2)$$

  where typical elements are $\Delta y_{nt} = y_{nt} - y_{n,t-1}, t = 2, ..., T$
  2. Fixed effects method (within transformation) (**remove Z**)
  3. Least Square dummy variables (LSDV) method (**account for Z**): include time-invariant individual effects by introducing them via individual intercept (N dummy variables $\mu_n, n = 1, ..., N$).
  - Degrees of freedom are $NT - N - K$ (impossible to estimate for small T).
  - $\hat{\beta}$ is $NT-$consistent, μ_n are $T-$consistent.

Fixed effects method

- LSDV consistency depends on N and T and may be numerically inefficient
- An equivalent formulation is to trasform the data by subtracting the averages by individual to every variable:

$$y_{nt} - \bar{y}_{n.} = (x_{nt} - \bar{x}_{n.})\beta + (\upsilon_{nt} - \bar{\upsilon}_{n.}) \tag{3}$$

  where $\bar{y}_{n.}$ and $\bar{x}_{n.}$ are the individual means of y and x
- α and μ_n disappears because they are time-invariant
- Fixed effect model is equivalent to estimate *LSDV*, but in *LSDV* μ_n are estimated directly, in fixed effects they are not (but it is possible to recover that).

Application

Application in R: example 1.1 (all solutions) and 1.2 (Croissant, Millo)

General notation

- P: probability, E: expected value, V: variance
- tr: trace of a matrix (sum of principal diagonal)
- cor: correlation, σ: standard deviation
- q: quadratic form
- I: identity matrix
- P: $X(X^TX)^{-1}X$, M: $1 - P$
- C: Cholesky matrix decomposition, such that $CAC^T = I$
- lnL: objective function of the maximum likelihood
- LR and LM are, respectively, the Likelihood ratio and the Lagrange multiplier

Index

- A panel is composed by N individuals denoted with n
- Each individual is observed during T time periods, denoted with t
- Sample size is O, with $O = NT$
- The K covariates are indexed by k (If we also have the column of one for the intercept, we have $K + 1$)

Two way error component model (i)

$$y_{nt} = \alpha + \beta^T x_{nt} + e_{nt} = \gamma^T z_{nt} + e_{nt}$$

$$\text{with } e_{nt} = \mu_n + \lambda_t + \upsilon_{nt}$$

- y is the response, α the intercept, x is the vector of K covariates with associated coefficients β, z, where $z_{nt}^T = (1, x_{nt}^T)$, γ: $\gamma^T = (\alpha, \beta^T)$
- e is the sum of time-invariant individual effect μ, individual-invariant time effect λ and the *iid* residual error υ
- The variance is σ^2, so σ_μ^2, σ_λ^2, σ_υ^2 and σ_e^2 are the variance of the four terms
- Estimated parameters are expressed with an hat ($\hat{\beta}$, $\hat{\sigma}^2$, etc...)

Two way error component model (ii)

- In matrix form:

$$y = \alpha j + X\beta + e = Z\gamma + e$$

$$e = D_\mu \mu + D_\lambda \lambda + \upsilon$$

where j is a vector of ones, X and Z the covariate matrix, μ the vector of N individual effects, λ the vector of T time effects, υ the vector of O residual effects

- D denotes a matrix of dummy variables
- Denoting by $J = jj^T$ a squared matrix of ones, we have

$$D_\mu = I_N \otimes J_T$$

$$D_\lambda = J_T \otimes I_N$$

- The covariance matrix (of e) is $\Omega_e = \sigma_\upsilon^2 I_{NT} + \sigma_\mu^2 I_N \otimes J_T + \sigma_\lambda^2 J_T \otimes I_N$

One way (error) component model - Transformation (i)

In this model the time-invariant terms disappear

- $S = I_N \otimes J_T$ is the matrix that, if post multiplied by a variable, returns a vector of length O containing the individual sums of the variables (each one repeated T times)
- $\bar{I} = I - \bar{J}$ is the matrix that, if post multiplied by a variable, returns the variable in deviation from its overall mean
- $B = \frac{1}{T}S$ and $W = I_{NT} - B$ are, respectively, the *between* and the *within* matrix
- $\Omega_e = \sigma_v^2(W + \frac{\sigma_1^2}{\sigma_v^2}B)$, where $\sigma_1^2 = \sigma_v^2 + T\sigma_\mu^2$. $\phi = \frac{\sigma_v^2}{\sigma_1^2}$
- $\theta = 1 - \phi$ is the fraction of the individual mean that is subtracted in the generalized least squared (GLS) model.

Two way error component model - Transformation (i)

In this model we have two different between matrices:

$$B_\mu = I_N \otimes J_T/T; B_\lambda = J_T \otimes I_N/N$$

- The within matrix $\bar{J} = J_{NT}/NT$, if post multiplied by a variable, returns the vector of the overall mean repeated NT times:

$$W = I - B_\mu - B_\lambda + \bar{J}$$

- $\Omega_e = \sigma_\upsilon^2(W + \frac{1}{\phi_\mu^2}\bar{B}_\mu + \frac{1}{\phi_\lambda^2}\bar{B}_\lambda + \frac{1}{\phi_2^2}\bar{J})$, where $\bar{B}_\mu = B_\mu - \bar{J}$, $\bar{B}_\lambda = B_\lambda - \bar{J}$, $\phi_\mu^2 = \frac{\sigma_\upsilon}{\sqrt{\sigma_\upsilon + T\sigma_\mu^2}}$, $\phi_\lambda^2 = \frac{\sigma_\upsilon}{\sqrt{\sigma_\upsilon + N\sigma_\lambda^2}}$, $\phi_2^2 = \frac{\sigma_\upsilon}{\sqrt{\sigma_\upsilon + T\sigma_\mu^2 + N\sigma_\lambda^2}}$
- $\theta_i = 1 - \phi_i, i = \mu, \lambda, 2$

References

- Mundlak, Y. (1961). Empirical production function free of management bias. Journal of Farm Economics, 43(1), 44-56.
- Baltagi, B. (2008). Econometric analysis of panel data. John Wiley & Sons.
- Baltagi, B. H., & Levin, D. (1992). Cigarette taxation: Raising revenues and reducing consumption. Structural Change and Economic Dynamics, 3(2), 321-335.
- Greene, W. H. (2000). Econometric analysis 4th edition. International edition, New Jersey: Prentice Hall, 201-215.
- Kalton, G., Kasprzyk, D., & McMillen, D. (1989). Panel surveys.
- Stock, J. H., & Watson, M. W. (2012). Introduction to econometrics (Vol. 3). New York: Pearson.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.