Quantitative Methods
Regression models and panel data
Part I: Introduction

BIg DAta MAnagement (BIDAMA) - Cycle XXVII

DIpartimento di Scienze Aziendali - Management e Innovation Systems (DISA-MIS)

Rodolfo Metulini

✉ rmetulini@unisa.it

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

Outline

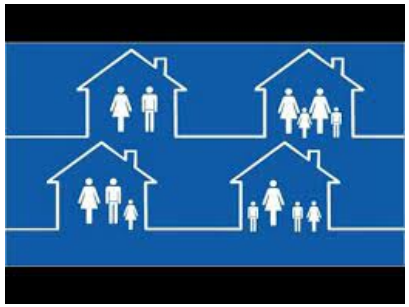
To begin with

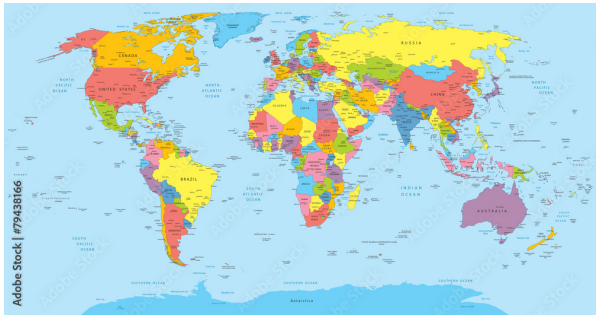
- Student's presentation.
- Presentation of myself.
- Presentation of the short course.
- Syllabus, textbook and material.
- Examination methods.

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 ans the 5 V's

Big data are characterized for their:

1. Volume
2. Velocity
3. Variety / Variability.

Big data needs to produce results that are able to:

1. Control for uncertainty (Veracity).
2. Create Value.

Panel data & Big data

- Panel data are characterized by their **large volume** (double indexed).
- Panel data analysis allows to **create value**, by mainly addressing the issue of heterogeneity along individuals and time.
- Heterogeneity is connected to many V's because:
  1. heterogeneity emerges with high volume of data,
  2. source of uncertainty (which is the other side of veracity) may increases when both individual and time variations are considered,
  3. Especially in financial analysis, data might be heterogeneous (may vary) at a very small time period extent (velocity),
  4. to produce the panel dataset it may be needed to extract data from many sources (variety),
  5. the value is increased, as modelling heterogeneity means that more complex information are retrieved from data.

What we do and we do not

We learn how to investigate the (causal) relation among features by means of panel data models, assuming that:

- the true relation between the variables is **linear**,
- the dependent variable (Y) is **quantitative**, so that we can assume a **normal** distribution for Y.

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

are **not the object** of this course.

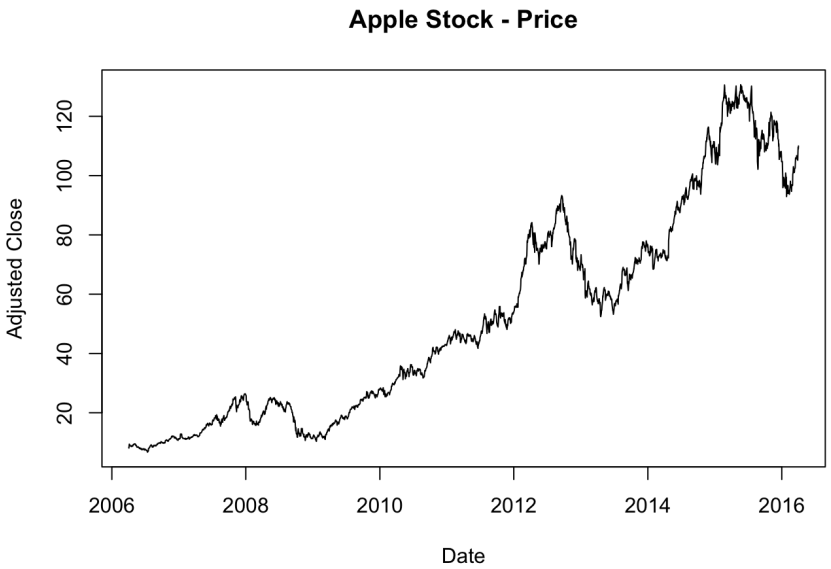
Many books addressing models for non linear relations and/or count data exists, but here we limit our attention to linear relations among (possibly) normal variables

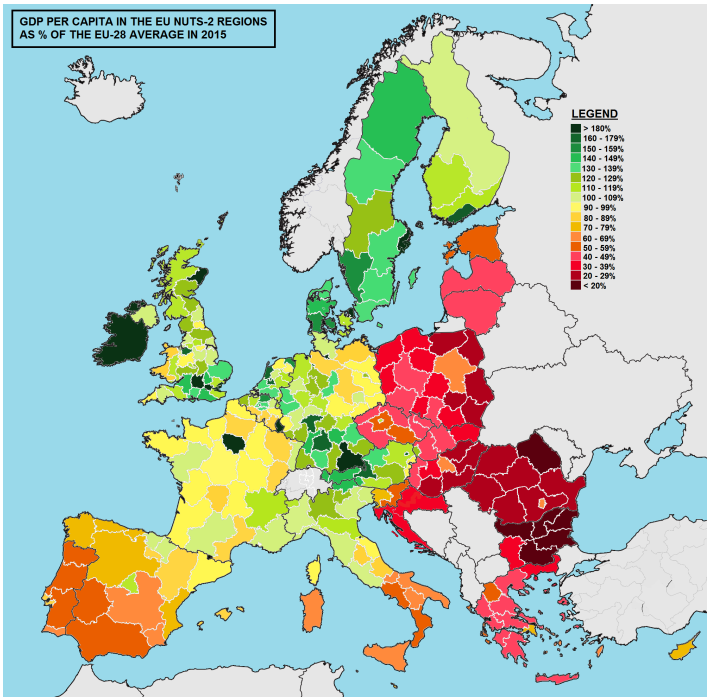
Cross sections and time series (i)

- In statistics and econometrics ... a **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.
- ... 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 study variation in time, with cross section variation inter-individuals (just at a limited extent).
- Examples of time series: i) real GDP by trimester, from Q1.2007 to Q4.2020. ii) Daily variation of Unicredit on Stock exchange market.
- Example of cross sections: i) the labour cost per capita (per hour, averaged over year 2019) for a sample of Chinese firms producing rice. ii) the GDP growth for NUTS2 regions from 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

... 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[1])

[1] This latter case is not an object of this short course

Benefits from using panel data

According to Baltagi (2008), panel data:

1. permit to control for individual heterogeneity;
2. are associated to 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 behavioural models than purely cross-section or time-series data.

1. Control 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

2. 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 $y_{nt} = \alpha + \beta_n x_{nt} + e_{nt}$

3. 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.

4. 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

5. 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

Can be classified in (Better to say "problems with data collection"):

- Design and data collection problems.
- Distortions of measurement errors.
- Selectivity problems:
  1. self-selectivity;
  2. non-response;
  3. attrition.
- Short time-series dimension.

Design and data collection

- These issues include:
  1. problems of coverage (incomplete account of the population of interest)
  2. non-response (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

- 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

- **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 $n * t$ asymptotical consistency relies on the number of individuals n tending to infinity.

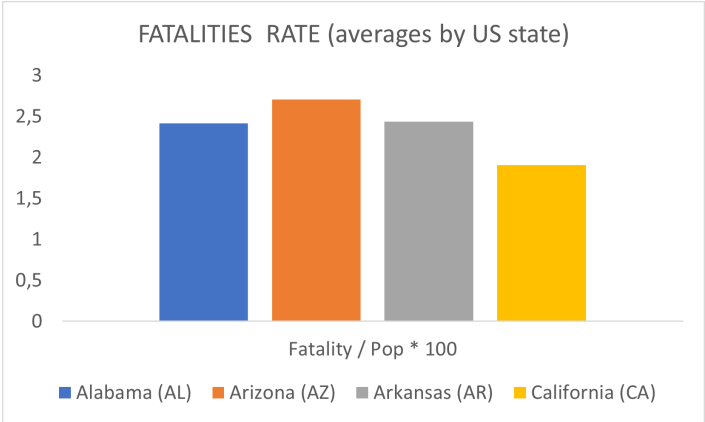
Individual heterogeneity: example (i)

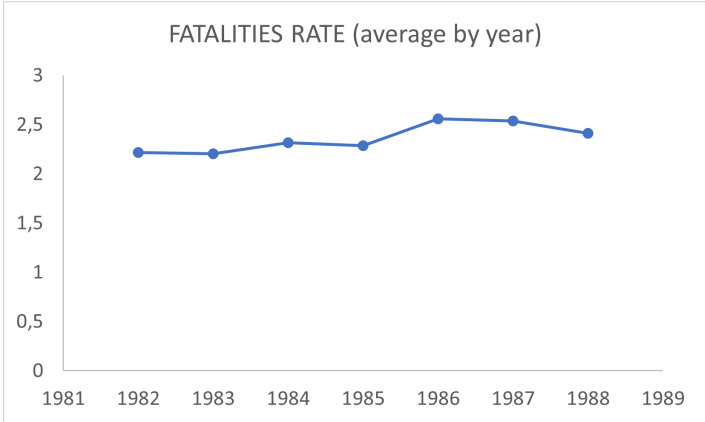
	state	year	beertax	frate
1	al	1982	1.53937948	2.12836
2	al	1983	1.78899074	2.34848
3	al	1984	1.71428561	2.33643
4	al	1985	1.65254235	2.19348
5	al	1986	1.60990703	2.66914
6	al	1987	1.55999994	2.71859
7	al	1988	1.50144362	2.49391
8	az	1982	0.21479714	2.49914
9	az	1983	0.20642203	2.26738
10	az	1984	0.29670331	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

Individual heterogeneity: example (ii)

- With panel we have more than one information per year and more than one information per state
- So, computing averages by state and averages by year is allowed.

  Figures highlight, respectively, the presence of inter-state and inter-year heterogeneity





- Each state presents a different level (heterogeneity among individuals).
- Each year presents a different level as well (heterogeneity among times).

Individual heterogeneity: example (iii)

The main interest when using regression models is that of correctly studying the causal relation between the dependent variable Y and a set of independents called X. In doing so, it is important to correctly account for individual heterogeneity in Y.

Stock and Watson (2007) 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

Individual heterogeneity: example (iv)

- Local beer taxation depends on (unobserved, not measurable) state-level characteristics (not included in the first two models).
- It is well proved in statistics that, if I miss to include something relevant in the model, and this is correlated with X, *OLS* returns biased and inconsistent estimated coefficients.
- 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 (e.g. cultural and religious people habits) are generally unknown.
- TAKE HOME MESSAGE: By using cross section data, the risk is to obtain wrong results on the causal effect of a regressor on the dependent.

Application

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

Statistical distributions (a non-exhaustive overview)

A statistical distribution is a function that assigns to any value x a value of probability $f(x)$. It worth recalling the main statistical distributions for continuous variables.

- **Normal (Gaussian)**: $N(\mu, \sigma^2)$, $-\infty < \mu < +\infty$, $\sigma > 0$

  $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$, for $-\infty < x < +\infty$

  $F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$, for $-\infty < x < +\infty$

  $\mathrm{E(X)} = \mu$, $\mathrm{V(X)} = \sigma^2$
- **Standardized Normal**: $\mathrm{Z} = \frac{X-\mu}{\sigma}$. $Z \sim N(0, 1)$. Support: [-∞, + ∞]. $\mathrm{E(Z)} = 0$, $\mathrm{V(Z)} = 1$.
- **T-student**: $T_n = \frac{Z}{\sqrt{\frac{k}{n}}}$ where n are degrees of freedom and $k \sim \chi^2$. Support: [-∞, + ∞].
- **Chi-squared**: $Q = \sum_{i=1}^{k} Z_i^2$; $Q \sim \chi_k^2$. Support: [0, +∞]. $\mathrm{E(Q)} = \mathrm{k}$. $\mathrm{V(Q)} = 2\mathrm{k}$.
- **F-Fisher**: $\mathrm{F} = \frac{\frac{X}{n}}{\frac{Y}{m}}$ where $X \sim \chi_n^2$ and $Y \sim \chi_m^2$. Support: [0, +∞]. $E(F) = \frac{n}{n-2}$

A stochastic approach

- We work with stochastic variables, or "variabili aleatorie" (v.a.)
- When considering a sample of observations (e.g. the vector of the annual income of n workers, the vector of the annual gross domestic product for n countries, etc..), statistically, we consider each value of the vector a realization from a stochastic variable, so...
- y_i (income of the $i-$th worker) is the stochastic realization of a v.a. (Y)
- Often we assume that Y is a $v.a.$ with a normal distribution, so $Y \sim N(\mu, \sigma)$, where $E(Y) = \mu$ and $V(Y) = \sigma^2$
- Probability density function: $f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}$
- Important: each $i - th$ observation is a realization of a stochastic variable. With notation $y_i \sim iid\ N(\mu, \sigma), i = 1, ..., n$ we mean that all the elements in the vector y are independent each others and share the same distribution.

Independence between two (or more) variables

- In linear regression models we work with two (or more) stochastic variables, let say, Y and X
- Recalling probability, two **events** A and B, if $P(A \cap B) = P(A)P(B)$, are not dependent ($A \perp\!\!\!\perp B$). Also $P(A \mid B) = P(A)$.
- When talking about stochastic **variables** (let say Y and X), we say that, if $E(Y \mid X) = E(Y)$, it follows that $X \perp\!\!\!\perp Y$.
- **Linear independence** between two variables can be measured on the vector of realizations x and y. If there exists scalars a_1, a_2 such that $a_1 x + a_2 y = 0$, where 0 is a vector of zeros, X and Y are dependent. The same can be generalized for more than two variables (e.g. $X_1, X_2, .., X_p$).
- A more elegant way to check for linear independence is to measure the rank of the design matrix of dimension nxp $X = [X_1, X_2, ..., X_p]$. If the rank is p, the p variables are each others independent.

Covariance and correlations

- A measure used for linear dependence among two variables is the covariance (and the correlation)

$$Cor(X, Y) = 0 (uncorrelation) \nrightarrow X \perp\!\!\!\perp Y$$

$$X \perp\!\!\!\perp Y \rightarrow Cor(X, Y) = 0$$

- $Cov(X, Y) = E(XY) - E(X)E(Y)$ (for X and Y two stochastic variables)
- $Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ (for sample realizations from X and Y)
- $Cor(X, Y) = \frac{Cov(X,Y)}{\sqrt{var(X)var(Y)}}$, where $var(X) = \frac{\sum_{i=1}^{n} (x_1 - \bar{x})^2}{n}$

The regression function (i)

- The most general formulation for the causal relation between one or a set of independent variables and a dependent variable can be expressed, **deterministically**, as:

$$Y = f(X_1, X_2, ..., X_p)$$

- If we just consider one independent variable:

$$Y = f(X)$$

- If we consider f to be a linear function, the most simple formulation to measure the linear relation is:

$$Y = \beta_o + \beta_1 * X$$

The regression function (ii)

- Switching to a "non deterministic" or "stochastic" formulation, the causal relation between Y and X may be expressed as:

$$Y = f(X) + \varepsilon$$

- ε (also called "disturbance") is a stochastic variable with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$ which permits to take into account in the model the effects of all not considered variables that may have an effect on Y.
- E.G. let consider a group of bank customers with the same income (where income is the variable X). It is unlikely that all of them will have exactly the same savings $Y = f(X)$.

The regression function (iii)

- The regression function is deterministic on X and "stochastic" on ϵ:

$$E(Y \mid X) = E(f(X) + \varepsilon) = E(f(X)) + E(\varepsilon) = f(X)$$

- Given a sample of realizations (empirical observations) $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, it is possible to explicit the regression function for each of the $i-$realization of the sample (generally, of dimension n):

$$y_i = f(x_i) + \varepsilon_i, \forall i = 1, 2, ..., n$$

$$\varepsilon_i \sim N(0, \sigma^2), \forall i = 1, 2, ..., n$$

Specification (i)

- The linear specification reads as:

$$y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i, \forall i = 1, 2, ..., n$$

with

$$E(Y \mid X) = \beta_0 + \beta_1 * x$$

- The expected value for the dependent variable y_i ($\hat{y}_i$) is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 * x_i$$

where

$$\epsilon_i = y_i - \hat{y}_i$$

is the sample realization of the stochastic variable ε on the $i-$th sample realization; $\hat{\beta}_0$ and $\hat{\beta}_1$ have to be estimated.

Model assumptions (i)

- 1) **Linearity on the parameters**: The relation between Y and X is linear

  Warning: the following formulations are linear in the parameters:

  $y_i = \beta_0 + \beta_1 * log(x_i) + \varepsilon_i$

  $y_i = \beta_0 + \beta_1 * \frac{1}{x_i} + \varepsilon_i$.

  This is not linear in the parameters:

  $y_i = \beta_0 + \beta_1^2 * x_i + \varepsilon_i$

- 2) **Zero mean of the disturbances**: $E(\varepsilon_i) = 0, \forall i = 1, ..., n$
- 3) **Homoschedasticity in the disturbances**: $Var(\varepsilon_i) = \sigma^2 < -\infty, \forall i = 1, ..., n$

Model assumptions (ii)

- 4) **Independence between the disturbances**: $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$
- 5) **Normality of the disturbances**: $\varepsilon_i \sim N(0, \sigma^2)$

  The error terms are normally distributed
- 6) **X deterministic**: The indipendent variable X is known without errors.

Ordinary Least Squares (i)

- The most used estimation method for determining an estimate for parameters β_0 and β_1 is that of Ordinary Least Squares (OLS).
- The objective is that of minimize a function, called Q:
- $\min_{\beta_0,\beta_1} Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 * x_i)^2$
- The solution of the problem will be:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$$

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Ordinary Least Squares (ii)

In summary:

- We denote with β_0 e β_1 the true (unknown) parameters expressing the causal relation between X and Y.
- We instead denote with $\hat{\beta}_0$ and $\hat{\beta}_1$ the corresponding OLS estimates based on the sample realizations of the variables X and Y.

Properties of OLS estimates (i)

- **Linearity**: $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the sample realizations $y_1, y_2, ..., y_n$ and $x_1, x_2, ..., x_n$
- **Unbiasness**:

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

- **Efficiency**: Among all estimators for β made with a linear combination of all the sample realizations, OLS estimates are those with smaller variance.

Properties of OLS estimates (ii)

- **Consistency**: For sample size n that tends to ∞, the estimates converge to the value that the estimator is designed to estimate (with 0 variance).
- **Best Linear Unbiased Estimators (BLUE)**: OLS estimates for β_0 and β_1 and those with smaller variance, among all the possible unbiased estimators.
- **Asymptotic normality**: OLS estimates for β_0 and β_1 distribute normally as n tends to ∞.

An unbiased estimator for the variance

- $SS_\varepsilon = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$ stays for "Sum of squares of errors" and it is an empirical measure of deviance of the disturbances.
- $\frac{SS_\varepsilon}{n-2}$ is the empirical variance.
- $\frac{SS_\varepsilon}{n-2}$ is an unbiased estimator for σ^2:

$$E(\frac{SS_\varepsilon}{n-2}) = \sigma^2$$

Variance decomposition

- The linear regression model admits the following variance decomposition:

$$Var_{total} = Var_{model} + Var_{residual}$$

or, deviance decomposition:

$$Dev_{total} = Dev_{model} + Dev_{residual}$$

which can be equivalently expressed as:

$$SS_T = SS_R + SS_\varepsilon$$

or:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The index of linear determination, R^2

- $R^2 = \frac{SS_R}{SS_T} = \frac{Var_{model}}{Var_{total}}$
  or, equivalently:

$$R^2 = 1 - \frac{SS_\varepsilon}{SS_T} = 1 - \frac{Var_{residual}}{Var_{total}}$$

- R^2 admits value in the range [0,1]
- $R^2 = 0$ when the model is completely **inadequate** to explain the relation among X and Y, according to the sample data;
  $R^2 = 1$ when the model is completely **adequate** to explain the relation among X and Y, according to the sample data.

The multiple linear regression model

- Let now assume the following relation is in place: $y_i = x_i'\beta + \varepsilon_i$
- y_i is the $i-$th sample realization from the stochastic variable Y.
- x_i is now a 1 x p **vector** containing the sample realizations of the p independent variables for the $i-$ realization: $x_i' = (x_{i,1}, ..., x_{i,p})$.
- β is the vector of regression coefficients of dimension p x 1.
- ε_i is the stochastic variable (scalar) for the disturbances.

The multiple linear regression model in matrix form (i)

- Let suppose to have n units with $n > p$.
- Let $\mathbf{Y} = (y_1, ..., y_n)'$ be the $nx1$ vector of the dependent variable for the n observations.
- Let $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$ be the $nx1$ vector of the disturbances for the n observations.
- Let $\mathbf{X}$ be the n x p matrix with the realizations of the n independent variables.

The multiple linear regression model in matrix form (ii)

- In a matricial formulation, the multiple linear regression model reads as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

- If the model contains the intercept, the first column of $\mathbf{X}$ is a vector of ones.

Assumptions of the multiple linear regression model

1. **X** is a deterministic matrix, possibly **with rank** p.
2. $\mathrm{E}(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.
3. $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.
4. $\varepsilon_i \sim N(0, \sigma^2)$.

We can summarise hypotesis 2 and 3 with:

$$E(\varepsilon\varepsilon^T) = \sigma^2 I_n$$

where I_n is an identity matrix of dimension n x n

Rank of a matrix (i)

- Let A be a rectangular matrix of dimension m x n.
- From A we can extract a number of squared submatrices, called "minors", by removing a row or a column.
- Let "order" be the number of columns (or rows) of that submatrix.
- For each minor we can compute the determinant.
- The rank of A is the largest order of the minors such that the determinant is different from 0.

Assumptions (discussion)

- If rank of **X** is p (assumption 1) it follows that all the independent variables are each others linearly independent (in other words, it does not exist an independent variable that can be written as a linear combination of the others).
- From assumption 2 it follows that:

$$E(y_i \mid x_i) = f(x_i), \forall i$$

and:

$$Var(y_i \mid x_i) = \sigma^2, \forall i$$

Estimates of β and σ (i)

- Let suppose assumptions 1–3 are in place.
- Let $\mathbf{Y}^T = (y_1, ..., y_n)$ be the 1 x n vector of the sample realizations from the v.c. $\mathbf{Y}$.
- We can obtain an estimation of the vector β with OLS:

$$S = \sum_{i=1}^{n} (y_i - x_i\beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

- After calculations:

$$S = \mathbf{Y}^T\mathbf{Y} - \mathbf{X}^T\beta^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^TX\beta =$$

$$\mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

Estimates of β and σ (ii)

- To obtain $\hat{\beta}$ we apply the partial derivates of S in terms of each β_i. In matrix notation:

$$\frac{\delta S}{\delta \beta} = -2\mathbf{X}^T\mathbf{Y} + \mathbf{X}^T\mathbf{X}\beta + \mathbf{X}^T\mathbf{X}\beta = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta$$

- By setting the previous equation to 0 and by solving for β we have that:

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\beta$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$$

Estimates of β and σ (iii)

- Let $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ the n x 1 vector of estimated y
- Let $\epsilon = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ the n x 1 vector of empirical disturbances.
- The estimated (unbiased) variance of the residuals is:

$$\hat{\sigma}^2 = \frac{\epsilon^T \epsilon}{n - p}$$

Estimators' properties

- Univariate properties holds in the multivariate case as well
- Moreover, by reminding that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$
- By further considering that $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, we have that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T(\mathbf{X}\beta + \varepsilon))$$

- It follows that $E(\hat{\beta}) = E[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T(\mathbf{X}\beta + \varepsilon)] = \mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\beta] + E[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T)\varepsilon)] = \beta + (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T)E(\varepsilon) = \beta$

- $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

Coefficients interpretation (i)

- Let suppose to have just two independent variables: X_1 and X_2.
- The relation among $\mathbf{Y}$ and the two independent variables is:

$$E(\mathbf{Y} \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- β_1 represents the (average) effect on $\mathbf{Y}$ of a unitary variation on X_1, by taking X_2 constant.
- β_2 represents the (average) effect on $\mathbf{Y}$ of a unitary variation on X_2, by taking X_1 constant.
- β_0 represents the expected value of $\mathbf{Y}$ when $X_1 = 0$ and $X_2 = 0$.

Coefficients interpretation (ii)

- Let suppose X_1 varies of an amount equal to δX_1 and that X_2 remains constant.
- In light of a variation of X_1, $\mathbf{Y}$ varies as well:

$$\mathbf{Y} + \delta\mathbf{Y} = \beta_0 + \beta_1(X_1 + \delta X_1) + \beta_2 X_2$$

- We obtain that :

$$\delta\mathbf{Y} = \beta_1 \delta X_1$$

from which it follows:

$$\beta_1 = \frac{\delta\mathbf{Y}}{\delta X_1}$$

Adjusted R^2

- By including one more variable to the regression model, the R^2 increases, even when the independent variable is not significant.
- Fo such a reason is preferable to use the adjusted R^2 in the multiple linear regression model:

$$R^2_{adj} = 1 - \frac{n-1}{n-p}\frac{SS_\varepsilon}{SS_T}$$

- The correction factor $\frac{n-1}{n-p}$ is always larger than 1, so $R^2_{adj} < R^2$.

The t-student test

- We want to test the null hypothesis that the $j-th$ β coefficient (β_j) is not significantly different from 0, under the assumption σ^2 is unknown.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

- The t-statistic $t = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\hat{\sigma}^2 v_{jj}}}$ is used, where v_{jj} is the $jj-th$ element of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$
- t, under the null hypothesis, is distributed as a t-student with $n-p$ degrees of freedom.
- If empirical t is larger than tables' t or smaller than minus tables' t, we reject the null hypothesis.

The F statistic (i)

- Let suppose we want to test the following system of hypothesis:

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_{p-1} = 0 \\ H_1 : \beta_j \neq 0 \text{ for at least one } j,\ j = 1, ..., p-1 \end{cases}$$

- In other words, under the null hypothesis, all the independent variables (except for the intercept) do not have any significant effect on $\mathbf{Y}$.

The F statistic (ii)

- Let $SS_{\varepsilon,restricted}$ the residual sum of squares from the model with just the intercept.
- $SS_{\varepsilon,unrestricted}$ the residual sum of squares from the model with all the independent variables.
- The F statistics is:

$$F = \frac{(SS_{\varepsilon,restricted} - SS_{\varepsilon,unrestricted})/(p-1)}{(SS_{\varepsilon,unrestricted})/(n-p)}$$

- The F statistic (F_{obs}) must be compared with the theoretical F (F_{th}) on the tables. If $F_{obs} > F_{th}$, we reject the null hypothesis.

Unobserved heterogeneity

- From a statistical point of view, panel data mainly addresses the issue of **unobserved heterogeneity** (i.e., controlling for it to avoid biased estimations).
- Let consider the model $y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \varepsilon_i, i = 1, ..., n$, where x_i is the i-th realization of an observable regressor distributed as a variable X and z_i is **unobservable** from a variable Z.
- The "feasible" model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, ..., n$ and may suffer for omitted variable bias: OLS $\hat{\beta}_1$ is consistent and unbiased only if Z is uncorrelated with either X or Y. [2]
- 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_1.

[2] A simple way to detect for the presence of unobserved heterogeneity is to look whether the distribution of estimated residuals satisfies OLS assumptions.

Panel data solution

- The panel data model aims to avoid the unobserved heterogeneity problem.
- Let consider

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

a panel linear regression equation for individual n at time t, where the residual part is the sum of μ_n (time-invariant, it represents the unobservable characteristics on the individuals) and υ_{nt} (the idiosyncratic term, which is time- and individual-variant). [3]

- the objective is to eliminate (wipes out) μ_n from the model (or to estimate it, in a way).

[3] From now on, we slightly change notation: n = 1, ..., N for individuals, t = 1, ..., T for times; α is used in place of β_0.

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. **??** in terms of observables only (or adequately accounting for them):
  1. Least Square Dummy Variables (LSDV) method include time-invariant individual effects by introducing them via individual intercepts (N dummy variables $\mu_n, n = 1, ..., N$ are going to be estimated). CONS:
     - Degrees of freedom are $NT - N - K$ (impossible to estimate for small T).
     - $\hat{\beta}$ is $NT-$consistent, μ_n are $T-$consistent.
  2. Differencing method (μ_n disappears, OLS turns out to be consistent):

$$\Delta y_{nt} = \Delta x_{nt}^T \beta + \Delta \mu_n + \Delta \upsilon_{nt} \tag{2}$$

     where typical elements are $\Delta y_{nt} = y_{nt} - y_{n,t-1}, t = 2, ..., T$
  3. Fixed effects method (within transformation) (μ_n disappears)

Fixed effects method

- LSDV consistency depends on N and T and may be numerically inefficient
- The fixed effect within transformation is an equivalent formulation based on transform the data by subtracting the averages by individual to every variable:

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

  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* (same estimated β), but in *LSDV* μ_n are estimated, in fixed effects they are not (but it is possible to recover that in a second step).
- It worth noting that *within* model disregards intra-individual variation!

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
- I: identity matrix
- $P = X(X^TX)^{-1}X^T$ returns the fitted values when post-multiplied by a vector
- $M = I - P$ returns the residuals when post-multiplied by a vector
- A panel is composed by N individuals denoted with n
- Each individual is observed at 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)

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

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

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

Two way error component model (ii)

- In matrix form (all individuals at all times, compactly):

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

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

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

- 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
$\Omega_e = \sigma^2_\upsilon I_{NT} + \sigma^2_\mu I_N \otimes J_T + \sigma^2_\lambda J_T \otimes I_N$

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* matrices
- $\Omega_e = \sigma_v^2(W + \frac{\sigma_1^2}{\sigma_v^2}B)$, 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. (2007). Introduction to econometrics (Vol. 1). New York: Pearson.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.