Applied Econometrics
Lecture 1

Giovanni Marin[1]

[1]Università di Urbino

Università di Urbino
PhD Programme in Global Studies
Spring 2018

Outline of this module

- Beyond OLS (very brief sketch)
- Regression and causality: sources of endogeneity
- Accounting for time-invariant unobserved components: the fixed effect estimator
- Miscellaneous tips for econometrics
- A step back: FAQs for empirical analysis (in social sciences)
- The benchmark case: experiments
- Using (simple) econometrics wisely to identify casual relationships
- Solving endogeneity: instrumental variable approach
- Econometrics for policy evaluation: difference-in-differences, (propensity score) matching and regression discontinuity design
- Non-causal (but useful) uses of econometrics

Teaching material

- Angrist JD, Pischke JS (2009) Mostly Harmless Econometrics. An Empiricist's Companion. Princeton University Press.
  - Chapters 1 to 5 (except 'technicalities')
- Becker SO (2009) Methods to Estimate Causal Effects. Theory and Applications
  - Available at `http://www.restore.ac.uk/Longitudinal/surveynetwork/documents/master_class_notes.pdf`
- Slides - available on my homepage
  `http://www.giovamarin.altervista.org`
- Other papers

Exam

- Take home exam
- Content ⇒ essay on a (reasoned) replication of a paper with STATA
- Timing ⇒ TBD...

In many cases OLS are not the best option

- If the dependent variable is not continuous
  - Count variable (non-negative) ⇒ poisson, negative binomial, zero inflated poisson and NB
  - Dummy variable ⇒ probit, logit
  - Ordinal variable ⇒ ordered probit, ordered logit
  - Categorical (non-ordinal) variable ⇒ multinomial logit, multinomial probit, conditional logit
- If the dependent variable is truncated or censored
  - Tobit model ⇒ censoring and truncation (corner solution)
  - Heckman model ⇒ incidental truncation
- Presence of endogeneity
  - Instrumental variable estimator(s)
- Presence of spatial lags and/or dependence
- Dynamic models (dynamic panel - Diff-GMM and Sys-GMM)
- ...

But still they are very 'powerful'

- Linearity $\Rightarrow$ constant marginal effect
  - $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \Rightarrow dy/dx = \hat{\beta} \ \forall i$
- Estimated coefficients can be promptly interpreted in sign, significance and magnitude of the effects
- As long as there are positive degrees of freedom ($n - k > 0$), it is possible to run the model $\Rightarrow$ with non-linear models (based on maximum likelihood estimation) it often happens that even when $n > k$ the estimator does not work
- Low computing requirements
- Very nice properties (e.g. incidental parameter issue)
- If used wisely, EXTREMELY useful to do descriptive analysis

Why does endogeneity represent a problem?

- In presence of endogeneity, your estimated βs (for any kind of estimator...) do not represent a correct estimation of the 'true' parameters of the relationship under investigation
- The interpretation of 'endogenous' relationships, though still of interest, should be done with extreme care (still better than tossing a coin...)
- In presence of endogeneity, you cannot claim that 'x causes y' (or 'x does not cause y') because endogeneity impairs the possibility to claim causality links
- A relationship can be very strong (and predictive) but may entail no causality link
- In formulas, endogeneity occurs when $E(y_i - \hat{y}_i | x_i) \neq 0$

$\Rightarrow$ Endogeneity will always be an issue when you will send a paper to a journal or you will present a paper at a conference or seminar...

Sources of endogeneity

There are three sources of endogeneity

- Measurement error in the independent variable
- Reverse causality (simultaneity)
- Omitted variables

$\Rightarrow$ A good instrumental variable (IV) solves all these problems $\Rightarrow$ Not so easy to find a good IV $\Rightarrow$ good econometrics can help

Measurement error in the independent variable

- Let assume that the true model is $y = \alpha + \beta x + \varepsilon$, with $E(\varepsilon|x) = 0$
- x is not directly measured, but only measured with some (random) error $\Rightarrow \tilde{x} = x + \eta$ (or $x = \tilde{x} - \eta$) with (for example) $\eta \sim N(0, 1) \Rightarrow E(x) = E(\tilde{x})$
- As you only observe $\tilde{x}$, what you actually estimate is $y = a + b\tilde{x} + \phi$ which corresponds to estimating $y = a + bx + (b\eta + \phi)$
- However, you would like to know the impact of x on y, not the impact of $\tilde{x}$
- The residual in the regression is now $\psi = b\eta + \phi \Rightarrow$ by construction, however, this residual is positively correlated with the true independent variable x as $x = \tilde{x} - \eta \Rightarrow$ the measurement error η is an omitted variable! (see next slides)
- When there is a correlation between the omitted variable and the variable of interest, the estimated coefficient for the latter is biased $\Rightarrow$ endogeneity!
- Measurement errors always lead to an attenuation bias (see example in STATA)

Reverse causality (simultaneity)

- Estimates are biased as long as the causality link that is investigated runs in both directions
  - x causes y
  - y causes x
- What does 'causes' means in this framework?
  - Whenever you run a regression you always need to have a theoretical model in mind
  - If in your theoretical model (mathematical or conceptual) you believe (or the referee/editor believes...) that you have reverse causality, you are in trouble
- Typical example: the demand (or supply) function
  - Quantity causes prices
  - Prices cause quantity

Reverse causality: demand and supply

- Demand depends on prices and another variable z (e.g. income) $\Rightarrow$ $q = \alpha + \beta p + \gamma z + \varepsilon$
- Supply depends on quantity and another variable w (e.g. technology) $\Rightarrow$ $p = \tau + \theta q + \phi w + \eta$
- Let us substitute the supply function into the demand function and rearrange

$$q = \alpha + \beta(\tau + \theta q + \phi w + \eta) + \gamma z + \varepsilon$$
$$q(1 - \beta\theta) = (\alpha + \beta\tau) + \beta\phi w + \beta\eta + \gamma z + \varepsilon$$
$$q = \pi_0 + \pi_1 w + \pi_2 z + \xi$$

- It is thus impossible to retrieve the 'structural' parameters

Omitted variable

- Endogeneity emerges as long as the residual is correlated with the explanatory variable $\Rightarrow E(\varepsilon|x) \neq 0$
- If the residual contains variables that are correlated both to the independent and dependent variables, the $E(\varepsilon|x) = 0$ is not satisfied
- VERY IMPORTANT $\Rightarrow$ the exogeneity assumption fails only if the omitted factors are correlated BOTH to the independent and dependent variables
  - If the omitted variable is just correlated with the independent variable, no problem
  - If the omitted variable is just correlated with the dependent variable, no problem $\Rightarrow$ omission will reduce the predicting power of the regression (R squared) but would not influence the estimated coefficient for the main independent variable of interest
  - Often, while running regressions you care about causality rather than about prediction!
- The problem could be solved by accounting for these omitted variables by means of proxy variables

Omitted variable: direction of the bias (I)

- Assume that you want estimate the impact of education (x_1) on wages (y) $\Rightarrow$ returns on education
- The innate ability (e.g. the IQ) of the worker, described by x_2, is positively correlated with the success in education but also positively correlated (for any level of education) with the wage obtained $\Rightarrow$ if two people with a college degree (same x_1) have different ability, the one with greater ability will have a greater wage
- Ideally, you would like to estimate the following empirical model:

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

- Your prior beliefs about β_1 and β_2 will be that both will be positive
- Unfortunately, you just observe x_1 while x_2 remains unobserved and enters the error term
- You will estimate something like:

$$y_i = \alpha + \tilde{\beta}_1 x_{1,i} + \eta_i$$

Omitted variable: direction of the bias (II)

- Due to the omission of a variable (ability) that is correlated both with the dependent variable (education) and the independent variable (wage), the estimate of the coefficient for education $\tilde{\beta}_i$ will be biased
- How biased?

Table: Direction of the bias of β_1 when we omit x_2

	$corr(x_1, x_2) > 0$	$corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Panel data

- Panel data ⇒ the 'individual' (firm, person, region, sector, etc) i is observed in multiple periods t
- Two dimensions, i and t ⇒ $\{y_{it}, x_{it}\}$
- If the panel is balanced (i.e. for all is, we observe the same ts), sample size will be $N \times T$
- With microdata, you will usually deal with datasets with $N > T$, while with macro data it is often the case that $T > N$
- The main advantage of panel data is that, by observing the same individuals for many years, it is possible to partial out for unobserved factors that are individual-specific
- Accounting for individual-specific unobserved factors may help solving endogeneity in two cases:
  - If endogeneity is due to measurement error that is individual-specific and time-invariant ⇒ $\tilde{x}_{it} = x_{it} + \eta_i$
  - If endogeneity is due to omitted variables (correlated with both x_{it}s and y_{it}) that is individual-specific and time-invariant (μ_i)

Linear model with panel data

- With panel data, the linear model reads as follows:

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$$

- This is the 'pooled' model
- As long as $E(\varepsilon_{it}|x_i) = 0$ (strict exogeneity) holds, the estimated $\hat{\beta}$ with OLS is unbiased
- The pooled model accounts for two different sources of variation:
  - Between individuals variation $\Rightarrow y_{i.} \equiv \sum_t y_{it}/T$
  - Within individuals variation $\Rightarrow y_{it} - y_{i.}$
  - Using the pooled model does not help to solve endogeneity concerns

Linear model with panel data

- Assume that we have a situation in with unobserved individual-specific factors and time-invariant factors (z_i) are correlated with both y_{it} and x_{it}
  - Omitted variable problems
  - Measurement errors
- Panel data may help you in accounting for these factors in a flexible way $\Rightarrow$ inclusion of a (dummy) variable that partials out for these unobserved factors
- You can thus estimate the following linear model (least squares dummy variables LSDV):

$$y_{it} = \alpha + \beta x_{it} + (\mu_i + \varepsilon_{it})$$

- As $\sum_i \mu_i = 1 \forall i$, then fixed effects ($\mu_i$) are perfectly collinear with the constant α
- Fixed effects μ_i allow for heterogeneous intercepts (for each individual) in your linear model

Linear model with panel data

- The Fixed Effect model can be estimated with simple OLS (adding to the pooled specification $N-1$ individual fixed effects μ_i)
- Adding the fixed effects μ_i is equivalent to:
  - Estimating the 'within transformation' of the model $\Rightarrow$ $(y_{it} - y_{i.}) = \alpha + \beta(x_{it} - x_{i.}) + \varepsilon_{it}$
  - Estimating the pooled model adding the individual-level means for the independent variables $x_{i.}$ (Mundlak)
- It is also very common (and strongly suggested) to add, together with individual-specific fixed effects, a series of year-specific (and individual-invariant) fixed effects (τ_t)

Limitations of the fixed effect model

- The Fixed Effect model only employs within-individuals variation ⇒ deviations from each individual's average
- No consideration of the variance across individuals
- It is not possible to identify variables that have no time variation
- If changes in time of the explanatory variables within individuals are measured with errors, then the estimated coefficients with a fixed effect model will be downward biased
- Possible solution (with endogeneity concerns...) ⇒ Random Effect Model
  - μ_i is assumed to be random and not fixed ⇒ it will be part of the error term
  - Optimal combination of the between- and within-individual variation (only within-individual in fixed effects)
  - Assumption ⇒ μ_i should not be correlated with x_{it} ⇒ very demanding...

Dummy variables

- Dummy variables are binary variables that only take two values (0 and 1)
- A categorical variable can be transformed into a set of dummy variables
  - Variable: education attainment (no high school, high school, college, postgraduate)
  - The four categories can be identified by dummy variables, one for each category
  - This also works for non-ordinal categorical variables (e.g. countries, sectors, regions)
- Assume that the variable x_i is a dummy variable and that you estimate the model $y_i = \alpha + \beta x_i + \varepsilon_i$
- The interpretation of $\hat{\beta}$ is the difference in the expected value of y_i between individuals for which $x_i = 1$ and individuals for which $x_i = 0$
  - $E(y_i|x_i = 0) = \alpha + \beta \times 0 = \alpha$
  - $E(y_i|x_i = 1) = \alpha + \beta \times 1 = \alpha + \beta$
  - $E(y_i|x_i = 1) - E(y_i|x_i = 0) = \alpha + \beta - \alpha = \beta$

Dummy variables

- Assume that by 'transforming' your categorical variable into a set of dummy variables you get three mutually exclusive dummy variables
  - $D1_i$, $D2_i$ and $D3_i$
  - $D1_i = 1 \Rightarrow D2_i = 0, D3_i = 0$
  - $D2_i = 1 \Rightarrow D1_i = 0, D3_i = 0$
  - $D3_i = 1 \Rightarrow D1_i = 0, D2_i = 0$
  - This implies that $D1_i + D2_i + D3_i = 1 \forall i$
- Estimate a model in which you add all your dummy variables $y_i = \alpha + \beta_1 D1_i + \beta_2 D2_i + \beta_3 D3i + \varepsilon_i$
- The constant α will be perfectly collinear with the three dummy variables $\Rightarrow$ the constant is actually the coefficient of a variable made out of ones $\Rightarrow \alpha \times 1$
- In order to keep the constant, one of the dummy variables should be omitted $\Rightarrow$ baseline category

Logarithms: very useful, very dangerous!!!

- Concept of logarithm $\Rightarrow x^y = z \Rightarrow \log_x z = y$
- Natural logarithm $\Rightarrow \log_e \Rightarrow \log_e e^x = ln e^x = 1$
- $ln(2x) - ln(x) = ln(2) \forall$ values of $x \Rightarrow$ scale-free!
- Many variables, once transformed in logarithm, are normally distributed
- If you estimate a regression such as $ln(y) = \alpha + \beta ln(x)$, then $\hat{\beta}$ will be the elasticity of y with respect to $x \Rightarrow$ a 1 percent increase in x gives rise to a $\hat{\beta}$ percent increase in $y \Rightarrow \frac{\partial y/y}{\partial x/x}$
- The logarithm is defined only for $x > 0 \Rightarrow$ also $\log(0)$ is not defined!!
- You CANNOT take the logarithm of a variable with zeros or negative values!!!
- And you CANNOT do anything like $\log(1 + x)$ or $\log(0.01 + x)$!!!

Interaction variables

- Assume that you want to estimate the impact of education on wages, but you believe (or want to test) whether it is constant or it changes according to age or gender $\Rightarrow (\partial wage/\partial educ|age = a)$
- This can be done in two different ways
  - Estimate a different regression for each different level of the age variable
  - Estimate the returns of education 'conditional' on age $\Rightarrow$ interaction variable!

Interaction variables

- Interaction variables can be expressed as follows:

$$Wage_i = \alpha + \beta Educ_i Age_i + \varepsilon$$

- $\partial Wage_i / \partial Educ_i = \hat{\beta} Age_i$
- This may be a bit too demanding as an assumption $\Rightarrow$ effect linear and zero with Age=0

$$Wage_i = \alpha + \delta Educ_i + \beta Educ_i Age_i + \varepsilon$$

- $\partial Wage_i / \partial Educ_i = \hat{\delta} + \hat{\beta} Age_i$

$$Wage_i = \alpha + \delta Educ_i + \beta Educ_i Age_i + \pi Age_i + \varepsilon$$

Interaction variables

- What if you want to estimate the differential effect of education on wage between male and female?

$$Wage_i = \alpha + \delta Educ_i + \beta Educ_i Female_i + \pi Female_i + \varepsilon$$

- $\partial Wage_i / \partial Educ_i = \hat{\delta} + \hat{\beta} Female_i$
- The wage return on education for male ($Female_i = 0$) is $\hat{\delta}$
- The wage return on education for female ($Female_i = 1$) is $\hat{\delta} + \hat{\beta} \times 1$

How to prepare and read a regression table

[see examples]

Misc of misc...

- Before running regression, explore the data
  - Distribution ⇒ extreme values? need to take the logarithm?
  - Bivariate relationships ⇒ scatterplots
  - Correlation matrix
  - Evaluate trends or check 'expected' relationships
  - Identify (possible) missing values
- Whenever you believe it is needed (more often than you imagine...), use weights for your regressions ⇒ Molise (Luxembourg) should not count as much as Veneto (Germany) in driving your results!
- If OLS works, then it is likely that other 'strange' would also work well. If OLS does not work, maybe results obtained with other strange estimators are wrong...