EMERGING MARKETS - Lecture 2: Methodology refresher

Maria Perrotta

April 4, 2013

STOCKHOLM INSTITUTE OF TRANSITION ECONOMICS

SITE

STOCKHOLM INSTITUTE OF
TRANSITION ECONOMICS

http://www.hhs.se/SITE/Pages/default.aspx

My contact: maria.perrotta@hhs.se

Aim of this class

There are many different types of questions, data and tools for investigating them.

Econometric work is more robust in some contexts than in others.

Aim of this lecture: overview of main potential issues, to judge how much confidence should be placed on the findings of a particular study.

What is Econometrics?

> *Experience has shown that each of the three view-points of statistics, economic theory, and mathematics, is a necessary, but not by itself sufficient, condition for a real understanding of the* **quantitative relations** *in modern economic life. It is the unification of all three that is powerful. This unification constitutes econometrics.*

Ragnar Frisch

- Relations in economic life
- Quantitative relations

Example

We wish to investigate whether international aid affects economic development.
We collect information on a sample of developing countries. For each country in the sample, we have observations on two observable variables at a given point in time:

- *growth* = annual GDP growth rate
- *aid* = amount of international aid received in a year

Example (cont.)

What relationship generated these sample data? We hypothesize that each population value of growth, denoted as $growth_i$, is generated by a population regression equation (or DGP) of the form:

$$growth_i = f(aid_i) + u_i$$

$u_i =$ an unobserved random error term representing all unknown and unmeasured factors that determine the individual value of $growth_i$.

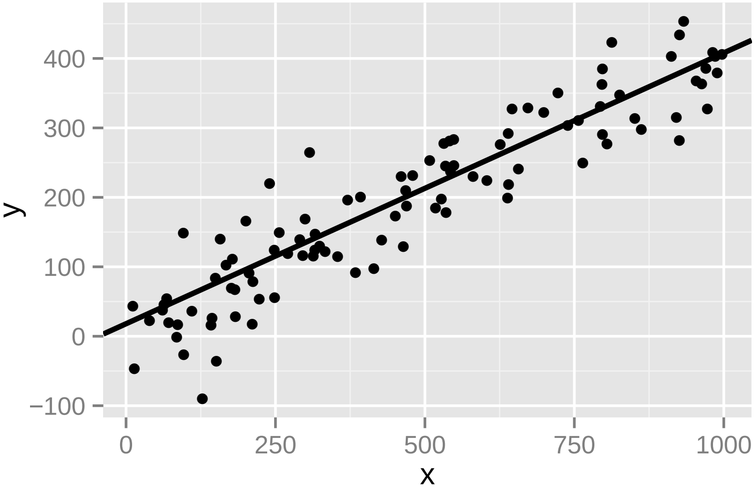
Example (cont.)

Typically, we hypothesize that the population regression function is a linear function.

$$growth_i = \beta_0 + \beta_1 aid_i + u_i$$

The regression coefficients β_0 and β_1 are unknown *parameters* characterizing the relationship. The first task of the analysis is to compute from sample data *estimates* of the regression coefficients.

Population regression function



Elements of econometric analysis

- **Data** Most economic data are observational, as distinct from data generated by an experiment. Cross-sectional, time-series, longitudinal and hierarchical data.
- **Specification** The econometric model that we think generated the data. Two aspects:
    - An economic model, derived from economic theory, or inspired by informal intuition and observation, specifies the main relation between the variables of interest: dependent, or outcome, and independent, or explanatory
    - A statistical model specifies the statistical properties (distribution, independence) of the variables in the relationship
- **Estimation** Choose the appropriate estimator: a function, a rule; compute estimates of the unknown parameters, using the sample data
- **Inference** Use the parameter estimates to test hypotheses about the population from which the sample was drawn, quantify uncertainty

Questions

- Descriptive analysis
- Causal relationships

Example

We observe that democratic countries tend to be richer. Can we claim that democratic institutions caused economic development?

Definitions of causality

Treatment framework

Fundamental problem of causal inference

Counterfactual questions: we never observe outcomes for the same individual under different treatment status at the same time

Identification strategy: defend the assumption that outcomes are independent of treatment status

FUQed questions

Methods

- Randomized experiment
- Natural experiment
- Observational studies

- Controlled regression studies
  - OLS
  - (PS) Matching
- Difference-in-difference
- Instrumental variables (IV)
- Structural model estimation
- Regression discontinuity

The ideal set-up

Controlled experiment: Isolate one factor, keep everything else constant, in a controlled environment

Randomized experiment: Second best, *black-box* definition of causality.
If the treatment has been randomly assigned, the outcome will be independent of treatment status. If outcomes are independent of treatment assignment, it is enough to compare averages between the two groups.

Problems with randomized experiments

- Costs: financial, ethical, “feasibility”
- Threats to internal validity: attrition, movements between groups
- Threats to external validity: duration, specificity, experimental effects, general equilibrium effects, scale effects

Deviations from the ideal

What is important to think about

- Observational data (observables and unobservables, measurement error, sizes, sample selection)
- Specification (estimates vs estimators, omitted variables, endogeneity)
- Interpretation of results (correlation vs causation, statistic and economic significance)

Matching

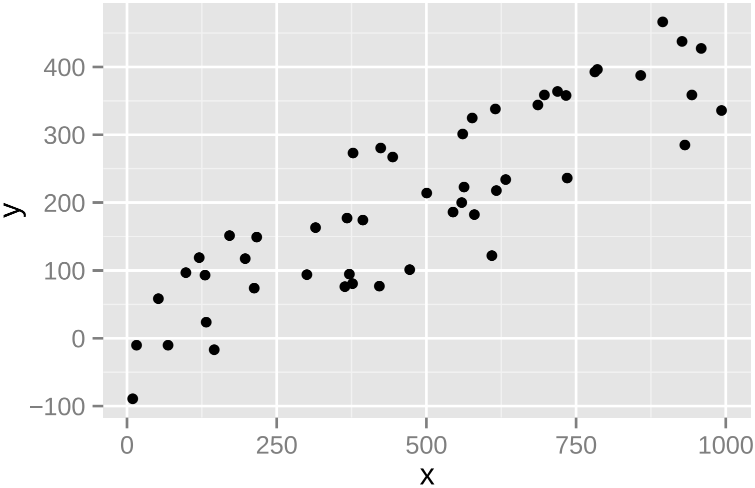
In some cases, the treatment is random *conditional* on a set of observable factors. If these factors X take discrete values (or can be discretized), we can compare treatment and control observations within each cell formed by the combination of the Xs.

Problems with matching

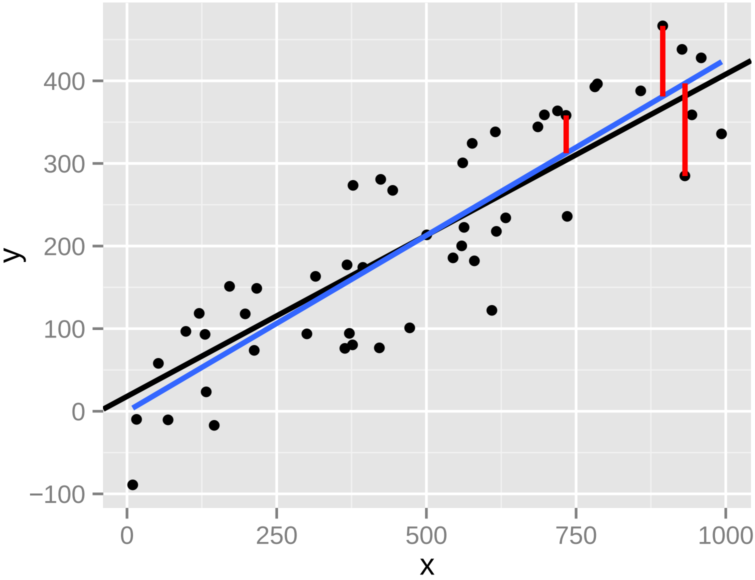
- Bias if treatment is not truly random conditional on observables (more bias than OLS)
- Small samples: cells containing only control or only treatment observations are dropped, loss of information $\Rightarrow$ propensity score matching
- Other problems common to regression approach

Controlled regression approach - OLS

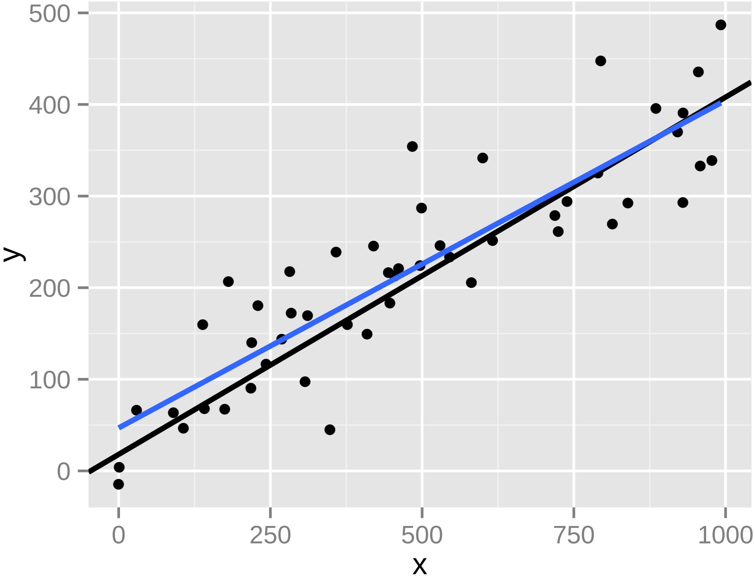
OLS is the basic regression design. Main intuition: How to draw a line through the data to estimate the (population) slope coefficient.



The idea behind OLS (ordinary least squares) is to minimize the predicted error terms.



Notice! The estimate $\hat{\beta}$ is a particular realization, depending on the sample



OLS - assumptions

- The population model is linear in parameters, as in

$$growth_i = \beta_0 + \beta_1 aid_i + u_i$$

- Observations are *i.i.d.* (random sample)
- There is variation in X
- $E(u_i|X_i) = 0$ *a.k.a.* strict exogeneity, zero conditional mean, or nonstochastic errors. Also implies that:
  - $E(u_i|X_i) = E(u_i) = 0$: treat X as fixed, there is no information in X that influence the error term.
  - $Cov(u_i, X_i) = 0$: the explanatory variable X is uncorrelated with the error term. (If not, then we say that X is an endogenous regressor.)
- Sferical errors (homoskedasticity and no serial correlation).
  - $Var(u_i) = \sigma^2$: the variance in u_i is constant, independent of the value of X_i. (If the variance differs between the different X, we have heteroskedasticity.)
  - $Cov(u_i, u_j) = 0$ for all i, j: no systematic relationship between two different error terms. (In some cases, it can be more realistic to assume a group structure, *clustered errors*.)

OLS - Relevant metrics

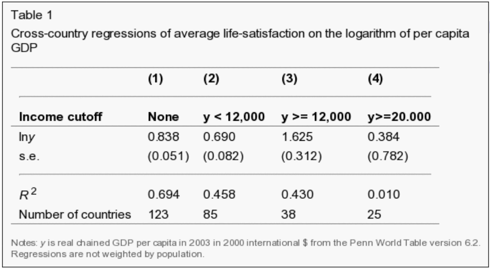


Table 1

Cross-country regressions of average life-satisfaction on the logarithm of per capita GDP

	(1)	**(2)**	**(3)**	**(4)**
Income cutoff	**None**	**y < 12,000**	**y >= 12,000**	**y>=20.000**
lny	0.838	0.690	1.625	0.384
s.e.	(0.051)	(0.082)	(0.312)	(0.782)
R^2	0.694	0.458	0.430	0.010
Number of countries	123	85	38	25

Notes: *y* is real chained GDP per capita in 2003 in 2000 international $ from the Penn World Table version 6.2. Regressions are not weighted by population.

- Coefficients (and their interpretation)
- Stars, statistical and economic significance
- Confidence intervals
- Goodness of fit. R^2 measures in percent how much of the variation in Y can be explained by the regression model. Caveats:
  - R^2 does not say anything about causality
  - Correlation≠causation (in bivariate regression, $R^2 = Corr(y,x)^2$)
  - Multivariate regression and explanatory power

Problems with OLS

- Omitted variable(s): $y \sim X$, omitting z.
  - Sign of the bias: $\hat{\beta} = \beta + \gamma * \delta$, where γ is the coefficient on $y \sim z$ and δ is the coefficient on $z \sim X$
  - Size of the bias. Typically there is no measure for the omitted variable. Compute how large the omitted variable bias must be to make results invalid (Altonji, Elder and Taber, 2000).

Example

Suppose X is binary. Compare:

$$\frac{E(\hat{y}|X = 1) - E(\hat{y}|X = 0)}{Var(\hat{y})}$$

with

$$\frac{E(u|X = 1) - E(u|X = 0)}{Var(u)}$$

Problems with OLS (cont.)

- Bad controls (caused by the variable of interest)
- Endogeneity (explanatory variable correlated with error term)
  - Reversed causation
  - Simultaneity
  - Sign of bias: $\hat{\beta} = \beta + \frac{Cov(X,u)}{Var(X)}$

Natural experiments and diff-in-diff

Sometimes the impact of X on Y can be studied through a change that affected X, for example a change in policy. These cases are called “natural experiments” because the change was not explicitly made in order to study the effect of the policy.

The obvious way to analyze these changes is to compare data from before and after, a *simple difference*.

Problem: How to distinguish the effect of the policy from a secular change?

- With only two periods, impossible
- Longer samples: look for a trend break BUT not for gradual reforms

Natural experiments and diff-in-diff (cont.)

To improve on the simple difference method, compare before and after the policy change for a group affected by the policy and a group not affected ($\sim$ Treatment and Control group).

Parallel trend assumption: This *difference-in-differences* gives an unbiased estimate of the policy effect if the average change in the outcome between before and after would have been the same in treatment and control group, in the absence of the policy.

Robustness checks with diff-in-diff

Diff-in-diff are very common. A number of checks can help establish if the analysis is convincing.

- Placebo: check whether there is an “effect” where there should not be
  - In another period, when no policy change happened
  - With a “fake” treatment group (i.e. not affected)
  - On a different outcome, that should not be affected
- Treatment and control group should be as similar as possible (in the case of randomized experiments, they are identical, at least in expectations) It can be useful to
  - Compare observable characteristics of the two groups
  - Interact those observables with the time dummy, to control for variations in the group composition and trends in observables
  - Use if possible several control groups, and compare the estimated effect

Problems with diff-in-diff

- Targeting based on differences, e.g. “Ashenfelter dip”
- Different average or starting level and functional form
- Long-run response vs. reliability
- Group structure - “clustered” standard errors, but need more than 2*2, and many clusters

Fixed effects

Regression with fixed effects is a generalization of diff-in-diff to many periods and many groups. Identification is obtained from within-group time variation. Useful in cases of self-selection into policy, or generally correlation between treatment asignment and outcome.

Advantage of many years and groups, more information. However, hard to check parallel trend assumption. Moreover fixed effects are not useful if the correlation depends on *trends* in outcomes, rather than levels.

Fixed effects also cause problems in models that should include lagged values of the explanatory variable (because the response takes more than one period) or the dependent variable (because of serial correlation) $\Rightarrow$ dynamic panel bias.

Instrumental variables

This approach is used for cases where X is correlated with u. An instrumental variable (IV) is a variable Z that

- is correlated with X
- is uncorrelated with u

Example: Acemoglu, Johnson and Robinson, 2001

Instrumental variables (cont.)

The estimation with IV is normally performed in two stages:

- The first stage $X \sim Z$ checks that X and Z are strongly correlated
- The second stage $Y \sim \hat{X}$. If Z is uncorrelated with Y beyond the effect *through* X (hence uncorrelated with ϵ, a.k.a. exogenous), $\hat{X}$, the part of X predicted by Z, is also exogenous and the coefficient is unbiased.

If there are more than one IV for each endogenous variable, the exogeneity can be tested, although the tests are not perfect.

Good and bad IV

Because it is difficult to test the validity of IV, they have to be convincing *a priori* on some good grounds. The best IV are generated by randomized or natural experiments. Examples include:

- Random encouragement designs
- Variables that approximate random encouragement design
- Natural experiment; the diff-in-diff estimation can be used as a first stage

Caveat: Even instruments that are randomly assigned can be invalid if they don't satisfy the exclusion restrictions, i.e. might affect the outcome directly

Problems with IV

- If the IV is not exogenous, the estimate is *more* biased than OLS
- Often the effect estimated cannot be interpreted to be a population average (LATE)
- Publication bias

Structural estimation

In many papers, authors write theoretical model that help causal inference (even generate IV).
This approach produces a complete framework, with theory and empirical tests, and estimates that are fully meaningful within the context of the model.
However, they are valid only to the extent that the model is valid and the assumptions realistic.

Ref. *Handbook of Econometrics*, Chapter 64

Regression discontinuity design

This approach is used when the treatment is a discontinuous function of an underlying continuous variable, for example

- Grameen bank eligibility rule, $< .5$ hectares
- Maimonides rule for class size, < 40 pupils
- IMF programs
- Election results

This type of rules imply that individuals with very close characteristics will (be exposed or not to the treatment and) end up in very different situations. Comparing individuals just below or just above the threshold makes the assignment as good as random.

Regression discontinuity design (cont.)

More realistically, the rule increases the probability to be treated:

- Rules are not always followed strictly
- Not everybody above the threshold will be treated, and somebody below it will.

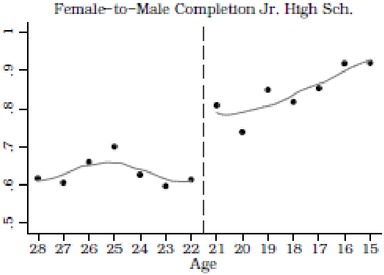
Focus on "intention to treat".

A possibility in these cases is to use an "IV version" of the RD estimator, with polynomial functions of the assignment variable.

Regression discontinuity design (cont.)

Important to think about:

- The data should show a discontinuity at the expected threshold



- Big samples are required, since only the observations around the threshold can be used.