INSEAD

The Business School
for the World®

PhD Programme
Jan/Feb 2016

Econometrics for PhDs

Professor: Amine Ouazad
amine.ouazad@insead.edu
Ext 5476 – Office 407D

Assistant: Endora Teo
endora.teo@insead.edu
Ext 5391 – Office 548 (Open office space)

Website: ***http://www.ouazad.com/*** *"Teaching Tab"*

Course Description

Who doesn't use data in daily life to justify one or more *hypothesis*? We talk effectiveness of medications, cost of driving a car, measures of behavior, returns of our savings, variance of currency exchange rates. This course is about sorting the wheat from the chaff, the factoids from the actual facts that the data reveals.

In particular, this course will be focused on finding *causal relationships* either from *observational* or *experimental* data. We will start by observing that most empirical questions can be framed as "*what is the (causal) impact of X on Y?*". In most situations, a correlation between X and Y is not indicative of a causal relationship between X and Y. For instance, if we see, across PhD classes, a positive correlation between your achievement and the achievement of your peers, is it: (1) a suggestion that your peers have an impact on your achievement? (2) a suggestion that you have an impact on your peers achievement? (3) a suggestion that some professors are better than others? (hence the group level correlation in achievement) or (4) a suggestion that students sort into classes by ability, either because of selection at enrollment or self-selection.

This central question of econometrics, the question of causality, is our approach to the standard tools of linear least squares, the Rubin causality model, panel data analysis, instrumental variable analysis, and natural experiments.

Materials

The following textbooks will be used. Mandatory textbooks are preceded by a *.

- *William H. Greene's *Econometric Analysis*.
- *A. Colin Cameron, Pravin K. Trivedi, *Microeconometrics Using Stata*, Stata Press, 2009.
- *J. Angrist and S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion,* Princeton University Press, 2009.
- A. Colin Cameron and P.K. Trivedi's Microeconometrics, Methods and Applications, at Cambridge University Press.

Everything else you need is either handed out in class or posted on the web.

Final examination

A 3-hour written examination at the end of econometrics A and at the end of econometrics B will be conducted. You will be asked questions of one of three kinds:

1. Comment of the analysis/results of the tables of a paper.
2. Econometric exercise of the same type as in W.H. Greene.
3. Questions on the implementation of a regression analysis in Stata.

Past exams are available either through me (amine.ouazad@insead.edu) or on the website.

Econometrics Project

Select a data set and an empirical question you would like to address. You need to find an interesting and feasible (data-wise) idea. Please consult with me and with your advisors. Send me the title of the empirical project about 3 weeks after the start of Econometrics A.

The final document will be at least 10 pages long, 12pt font size, margins 2cm, single space between lines. Structure it like a publishable paper: “Introduction,” “Data Set and Descriptive Evidence,” “Specification and Results,” “Discussion,” “Conclusion.”

When reading the text, I should find answers to the following questions (not necessarily in this order, but within each part).

1. **“Introduction”** What is the causal relationship of interest? Why do we care?

A causal relationship is useful for making predictions about the consequences of changing circumstances or policies; it tells us what would happen in alternative (or counterfactual) worlds.

2. **“Introduction”** What experiment could ideally be used to capture the causal effect of interest?

The gold standard is usually a random experiment. Research questions that cannot be answered by any experiment are FUQd: Fundamentally Unidentified Questions.

3. **“Data Set”** What is the data set used? Does it have interesting features that are helpful for identifying the causal relationship of interest? Does the data set have any drawbacks/flaws (lack of representativeness, attrition, top coding or missing values)?

4. **“Descriptive Evidence”** What is the descriptive evidence (correlations) that suggests there may be a relationship between the variables of interest? Why is such descriptive evidence potentially confounded by a number of other mechanisms (correlation is not causation)?

5. **“Specification”** What is the identification strategy? *An identification strategy is the manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment.* Are there issues with endogeneity of the explanatory variables, which are not discussed in the paper? Can you think why the dependent variable might be used to explain variation in the regressors? Are there other influences on the dependent variable that are not included in the regression? Do you think that they are correlated with the regressors? If so, what is the anticipated direction of the bias? Suggest other potential sources of bias. For each source of bias, write the data generating process, the estimating equation and the estimator. Explain why the estimator is biased.

6. In **“Specification”** What is the mode of statistical inference?

The answer to this question describes the assumptions made when constructing standard errors. Are there reasons to believe that the errors might be heteroskedastic or autocorrelated?

7. “**Discussion**” What additional robustness checks can be conducted?

8. “**Conclusion**” Why do we care about your results? Do they have policy or managerial implications?

Econometrics is all about stating the identification (and inference) assumptions. Identification is always obtained at the cost of assumptions, so the point of your paper is to make a convincing case that you have identified the relationship of interest with not-so-stringent assumptions.

Finally, replicability is key to modern research. Other researchers should be able to find the same results as your paper. Hence, provide in a single zip file: the pdf paper, the Stata do files that produces your tables/figures, and the Stata data set used for the analysis.

The paper is handed at the end of Econometrics A. Date announced in class.

Software

We will use Stata 13 or above in the course. Stata has the advantage of providing the right statistics for a paper – thus it is a great tool to learn parts of econometrics on the go.

Of course, in my research I use R, Python, Matlab, Fortran, and C, but these are tools that fit very specific ‘niche’ needs (e.g. Fortran for fast calculations, Matlab for optimization) and fall outside the scope of this course.[1]

[1] Stata is so nice and user-friendly. Note for instance that to obtain regression output in R takes two command lines (!). First, `regressionoutput<-lm(y~x)` and then `summary(regressionoutput)`. Obtaining

Frequently Asked Questions

I don't have a very strong mathematics background. Will I succeed?

- Relax. Both the discussion and the algebra are important in this class. In fact, focusing too much on the formal aspect (the algebra) can be a hindrance as econometrics is so much more than statistics. Of course, there will be derivations but we will always stop and ask us: is it a meaningful research approach? Can my data speak? What papers have used a similar approach? This is a conceptual course. In a number of years, I have seen a large number of students perform well with little prior maths background.

Shouldn't we just learn a lot of Stata?

- This refers to the previous question – before starting your data analysis using a statistical software package, make sure you have chosen your empirical question well, and that you ask yourself the right questions: (i) what is my causal effect of interest? (ii) under what conditions can I identify the causal effect of interest? (iii) do I need to add more variables, more observations? Sometimes little coding is necessary to write a good paper. You need to know the right approach, which is the focus of the in-class sessions. I will also guide you on how to use Stata in class.

clustered-corrected standard errors (see session 5) requires a long procedure in R. All of this is done in one simple line in Stata.

Session 1: Introduction & Identification

- **Key concepts:**
  - The identification problem.
  - Randomized experiments as the Golden Benchmark.
  - Asymptotic Statistics: Convergence, The Central Limit Theorem.

- **Chapters**
  - *Econometric Analysis*. Appendix A, B, C, D.
  - *Microeconometrics Using Stata*. Chapter 1. Ex 2,3,4,8.

- **Readings**
  - Marianne Bertrand and Simeon Djankov and Rema Hanna and Sendhil Mullainathan, 2007, *"Obtaining a Driver?s License in India: An Experimental Approach to Studying Corruption*,? The Quarterly Journal of Economics, MIT Press, vol. 122(4), pages 1639-1676, November.

Charles Manski calls it the "*fundamental problem of econometrics*." We called it the problem of finding *causal relationship* in the description of this course (above). Another way to put it is to ask "What would have happened if X had not happened?", e.g. what would have happened had you not used a child safety seat? Would your child have escaped unharmed from the accident? This is a tough question, because we do not observe the counterfactual, i.e. the situation of your child had he not been in a safety seat.

The problem of *the lack of observability of the counterfactual* is pervasive in management, TOM, OB, Finance, Marketing, Strategy, Entrepreneurship, and many other fields. Would firms perform better were they to hire more female directors? (question asked by one of your former PhD colleague). Would companies pay their CEOs less with a policy of Say on Pay ? (question asked by Prof. Maria Guadalupe).

We will see how to solve this problem, and when this issue is solved we will say that we have **identified the causal impact** of X on Y.

Session 2: Inference and Asymptotics

- **Key concepts:**
  - Convergence concepts.
  - The Law of Large Numbers.
  - The Central Limit Theorem.

- **Chapters**
  - *Econometric Analysis*. Appendix A, B, C, D.
  - *Microeconometrics Using Stata*. Chapter 1. Ex 2,3,4,8.

The problem of identification (previous session) is the first fundamental problem of econometrics. The second problem of econometrics is the problem of inference, i.e. the problems that arise because of a small number of observations.

Session 3: Linear Regression

- **Key concepts**: Ordinary Least Squares (OLS).
  - Properties: Consistence and Asymptotic Normality, Best Linear Unbiased Estimator.
  - Inference: confidence intervals, heteroskedasticity and clustering.
  - Tests: F-tests.
  - The Frisch-Waugh-Lovell theorem.

- **Chapters**
  - *Econometric Analysis*. Chapters 1, 2, 3, 4, 5, 8. Exercise 3.1, 3.6, 3.7, 4.7, 4.13, 5.3, 5.6, 8.1
  - *Microeconometrics using Stata*. Chapter 3, sections 1,2,3,4,8. Exercises 1, 2, 6. Chapter 5 (5.1, 5.2, 5.3).

- **Readings**
  - Steven D. Levitt, 2008, *"Evidence that Seat Belts Are as E?ective as Child Safety Seats in Preventing Death for Children Aged Two and Up*,? The Review of Economics and Statistics, MIT Press, vol. 90(1), pages 158-163, 07.

If you haven't seen **linear regression** in the past, this will become your Swiss army knife of data analysis. If you have seen linear regression in the past, you will leave this course realizing how important the framework is. We will see how linear regression tackles the problem of **identification** (the first fundamental problem of econometrics) and the problem of **inference** (the second fundamental problem of econometrics).

Session 4: Identification Issues in Linear Regressions

- **Key concepts**
  - Omitted variable bias.
  - Measurement error bias.
  - Functional form misspecification.

- **Chapters**
  - *Econometric Analysis*. Chapters 6, 7. Exercises 6.1, 6.2.
  - *Microeconometrics using Stata*. Chapter 3, sections 5, 7.

Linear regression allows us to identify causal effects under a set of assumptions (seen in the previous session). When the assumption of the exogeneity of the covariates is not satisfied, identification issues arise. This is the focus of this session. We will see three sources of violations of such assumptions: the omitted variable bias, the measurement error bias, and functional form misspecification. We will also provide ways to overcome these three issues.

Session 5: Inference Issues in Linear Regressions

- **Key concepts**
    - Heteroscedasticity
    - Clustering
    - Autocorrelation
- **Reading**
    - B.R. Moulton, “An illustration of a pitfall in estimating the effects of aggregate variables on micro units,” The Review of Economics and Statistics, 1990.

The linear regression framework (our Swiss army knife of data analysis) solves inferential issues (the second problem of econometrics) under the assumption of homskedasticity (spelling can differ according to the textbook). We will see what to do when we such assumption is not satisfied. Inferential issues are important – no paper can be written without proper management of inferential issues –, but are very rarely the spotlight of an applied research paper.

Session 6: Identification with Simultaneous Equations

- **Key concepts**
    - Conditions for identification in Simultaneous Equation Models.
- **Chapters**
    - *Econometric Analysis*. Chapter 13. Exercise 13.1.
    - *Microeconometrics using Stata*. No chapter.
- **Readings**
    - Sacerdote, Bruce. *“Peer Effects With Random Assignment: Results For Dartmouth Roommates*,? Quarterly Journal of Economics, 2001, v116(2,May), 681-704.

Solving the problem of identification is particularly difficult when looking at social relationships: what is the effect of my friends on my performance at the PhD program? Because we choose our friends endogenously, it is hard to obtain random (thus exogenous) variation in our peer group. This session addresses such issue.

In fields that make use of markets, a typical issue is the identification of demand and supply shocks: what caused the recent oil price crash? A downward demand shock or an upward supply shock? This problem raises similar questions as the previous problem of identifying causality in social groups.

This is the point at which we will need to introduction an extension of the linear regression model: the instrumental variable model, also called IV (pronounce aye-veeh).

Session 7: Instrumental variable estimation

- **Key concepts**
  - IV estimator, exclusion restriction
  - Consistency of the IV estimator
  - Bias of the IV estimator
  - Hausman test
- **Chapters**
  - *Econometric Analysis*. Chapter 12. Exercise 12.1, 12.5
  - *Microeconometrics using Stata*. Chapter 6 (sections 1 to 3).
- **Readings**
  - Acemoglu, Daron and Johnson, Simon and Robinson, James and Thaicharoen, Yunyong, 2003. *"Institutional causes, macroeconomic symptoms: volatility, crises and growth*," Journal of Monetary Economics, Elsevier, vol. 50(1), pages 49-123, January.

In the previous sessions, we saw four cases in which the assumption of the exogeneity of the covariates is not satisfied: omitted variable bias, measurement error, functional form misspecification, and the simultaneity problem. The first and the last issues (omitted variable bias, and simultaneity problem) cannot easily be tackled using the linear regression model.

This session introduces the instrumental variable model. Since the early 1990s, this approach to data analysis has become ubiquitous as it clearly spells outs conditions under which a causal effect is identified.

Session 8: Finding Good Instrumental Variables: Identification strategies

- **Key concepts**
  - Difference-in-differences.
  - Regression discontinuity design.

- **Chapters**
  - *Econometric Analysis*. No chapter.
  - *Microeconometrics using Stata*. No chapter.
  - *Mostly Harmless Econometrics*. Chapter 6, Getting a Little Jumpy: Regression Discontinuity Designs.

- **Readings**
  - David Card and Alan B. Krueger*, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania",* The American Economic Review, Vol. 84, No. 4 (Sep., 1994), pp. 772-793.
  - John J. Donohue and Steven D. Levitt, 2001. *"The Impact Of Legalized Abortion On Crime"*, The Quarterly Journal of Economics, MIT Press, vol. 116(2), pages 379-420, May.
  - John DiNardo, David S. Lee, "*Economic Impacts of New Unionization on Private Sector Employers: 1984-2001"*, Quarterly Journal of Economics, November 2004, Vol. 119, No. 4, Pages 1383-1441.

The paper by Acemoglu and Robinson in the previous session found an Instrumental Variable to identify the impact of institutions on economic development. But how do we find such instrumental variables? It turns out that two methods have been very successful strategies to approach the problem of finding instrumental variables: difference-in-differences and the RD design estimator (regression discontinuity design).

Session 9: Panel data estimation: The Longitudinal Dimension as the Exogenous Source of Identification

- Key concepts
  - random effects, fixed effects
  - within estimator, between estimator
  - dynamic panel data: bias, estimation using GMM
- Chapters
  - *Econometric Analysis*. Chapter 9. Exercise 9.1.
  - *Microeconometrics using Stata*. Chapter 8 (sections 8.1 to 8.9).
- Readings
  - Marianne Bertrand and Antoinette Schoar, 2003, *“Managing With Style: The Effect Of Managers On Firm Policies”*, The Quarterly Journal of Economics, MIT Press, vol. 118(4), pages 1169-1208, November.
  - Francine Lafontaine and Kathryn L. Shaw, 1999, *“The Dynamics of Franchise Contracting: Evidence from Panel Data*,? Journal of Political Economy, University of Chicago Press, vol. 107(5), pages 1041-1080, October.

With longitudinal data, i.e. data that follows firms, individuals, countries, regions, stocks, etc. over time, it is typically much easier to deal with the problem of causality.

Session 10: Bootstrapping for Confidence Intervals

- **Key concepts**
  - empirical c.d.f.
  - asymptotic refinement

- **Chapters**
  - *Econometric Analysis*. Section 17.6
  - *Microeconometrics using Stata*. Chapter 13 (Sections 1 to 4)

- **Readings**
  - R Stine, *“An introduction to bootstrap methods*,” Sociological Methods and Research, 1989.
  - Marianne Bertrand and Esther Dufo and Sendhil Mullainathan, 2004. *“How Much Should We Trust Difference-in-Differences Estimates”*, The Quarterly Journal of Economics, MIT Press, vol. 119(1), pages 249-275, February.

In session 5 we saw standard methods to deal with the problem of inference. However, the methods that we outlined sometimes require strong assumptions and intensive calculations. In this session, we introduce the bootstrap method for standard errors.