LISER

LUXEMBOURG INSTITUTE OF
SOCIO-ECONOMIC RESEARCH

INTERNATIONAL WORKSHOP
ON
CAUSAL INFERENCE, PROGRAM EVALUATION, AND EXTERNAL VALIDITY

JULY 13 · 14 2017

ESCH/ALZETTE · GRAND-DUCHY OF LUXEMBOURG

FONDS SOCIAL EUROPÉEN LUXEMBOURG

Avec le soutien du
Fonds social européen

THE GOVERNMENT
OF THE GRAND DUCHY OF LUXEMBOURG
Ministry of Labour, Employment and
the Social and Solidarity Economy

UNIVERSITÀ
DEGLI STUDI
FIRENZE
DISIA
DIPARTIMENTO DI
STATISTICA, INFORMATICA,
APPLICAZIONI "G.PARENTI"

arco
ACTION RESEARCH
FOR CO-DEVELOPMENT

Presentation of the Keynote speakers

Rajeev **Dehejia**

Rajeev Dehejia received his Ph.D. in economics from Harvard University in 1997. He has been on the faculty of the Department of Economics and The Fletcher School at Tufts University and of the Department of Economics and the School of International and Public Affairs at Columbia University, and has held visiting positions at Harvard, Princeton, and the London School of Economics. Rajeev's research spans econometrics, development economics, labor economics, and public economics, with a focus on empirical microeconomic policy research. His research interests include: econometric methods for program evaluation, financial development and growth, financial incentives and fertility decisions, moral hazard and automobile insurance, religion and consumption insurance, and the causes and consequences of child labor.

Rajeev's articles have appeared in The Journal of Law and Economics, The Journal of Human Resources, The Review of Economics and Statistics, the Journal of Business and Economic Statistics, the Journal of the American Statistical Association, The Quarterly Journal of Economics, the Journal of Econometrics, the Journal of Public Economics, the Journal of Development Economics, and Economic Development and Cultural Change. Rajeev is a Research Associate of the National Bureau of Economic Research, a Research Fellow at the Institut zur Zukunft der Arbeit (IZA), and a Research Network Fellow at CESifo. He is a coeditor of the Journal of Business and Economic Statistics.

Donald B. **Rubin**

Donald B. Rubin is John L. Loeb Professor of Statistics, Harvard University, where he has been professor since 1983, and Department Chair for 13 of those years. He has been elected to be a Fellow/Member/Honorary Member of: the Woodrow Wilson Society, John Simon Guggenheim Memorial Foundation, Alexander von Humbolt Foundation, American Statistical Association, Institute of Mathematical Statistics, International Statistical Institute, American Association for the Advancement of Science, American Academy of Arts and Sciences, European Association of Methodology, British Academy, and the U.S. National Academy of Sciences. He has authored/coauthored nearly 400 publications (including ten books), has four joint patents, and has made important contributions to statistical theory and methodology, particularly in causal inference, design and analysis of experiments and sample surveys, treatment of missing data, and Bayesian data analysis. Among his other awards and honors, Professor Rubin has received the Samuel S. Wilks Medal from the American Statistical Association, the Parzen Prize for Statistical Innovation, the Fisher Lectureship, and the George W. Snedecor Award of the Committee of Presidents of Statistical Societies. He was named Statistician of the Year, American Statistical Association, Boston and Chicago Chapters. He has served on the editorial boards of many journals, including: Journal of Educational Statistics, Journal of American Statistical Association, Biometrika, Survey Methodology, and Statistica Sinica. Professor Rubin has been, for many years, one of the most highly cited authors in mathematics in the world (ISI Science Watch), as well as in economics (Highly Cited Economists), with nearly 160,000 citations by autumn 2014, with over 16,000 in 2013 (according to Google Scholar). For decades he has given keynote lectures and short courses in the Americas, Europe, and Asia. He has also received honorary doctorate degrees from Otto Friedrich University, Bamberg, Germany; the University of Ljubljana, Slovenia, and Universidad Santo TomÃ¡s, BogotÃ¡, Colombia; as well as honorary professorships from University of Utrecht, The Netherlands; Nanjing University of Science & Technology, Nanjing, China; Xian University of Technology, Xian, China; and Shanghai Finance University, Shanghai, China.

Thursday, 13th July 2017

08:45-09:30 Welcome Coffee and participants' registration

09:30
10:30

Bridging Observational Studies and Randomized Experiments by Embedding the Former in the Latter

Keynote speech by Donald B. Rubin - Harvard University

Consider a statistical analysis that draws causal inferences using an observational data set, inferences that are presented as being valid in the standard frequentist senses; that is an analysis that produces (a) point estimates, which are presented as being approximately unbiased for their causal estimands, (b) p-values, which are presented as being valid in the sense of rejecting true causal null hypotheses at the nominal level or less often, and/or (c) confidence intervals, which are presented as having at least their nominal coverage for their causal estimands. For the hypothetical validity of these statements (that is, if certain explicit assumptions were true, then the validity of the statements would follow), the analysis must embed the observational study in a hypothetical randomized experiment that created the observed data set, or a subset of that data set. This effort is a multistage effort with thought-provoking tasks, especially in the first stage, which is purely conceptual. Other stages may often rely on modern computing to implement efficiently, but the first stage demands careful scientific argumentation to make the embedding plausible to scientific readers of the proffered statistical analysis. Otherwise, the resulting analysis is vulnerable to criticism for being simply a presentation of scientifically meaningless arithmetic calculations. Sadly, this perspective is rarely implemented with any rigor, for example, completely eschewing the first stage. Instead, often analyses appear to be conducted using computer programs run with limited understanding of or assessment of the assumptions of the methods being used, producing tables of numbers with recondite interpretations, but presented using jargon, which may be familiar but also may be scientifically impenetrable. These points will be illustrated using the analysis of an observational data set addressing the causal effects of parental smoking on their children's lung function. The conceptually most demanding tasks are often the most scientifically interesting to the dedicated researcher and readers of the resulting statistical analyses.

10:30-11:00 Coffee Break

11:00-12:30 FIRST SESSION: Applications of Program Evaluation

11:00
11:30

School Accountability, Score Manipulation and Economic Geography

Lorenzo Neri - Queen Mary University

We document how grading standards for exams at the end of primary education in England have triggered inflation of school quality indicators in national league tables. The cumulated effects over time resulted in significant differences in the quality signalled to parents for otherwise identical primary schools of the country. Institutional features ensure that these differences are as good as random, and reveal that inflation followed from discretion in grading of randomly assigned external markers. We use census data and administrative records on standardized tests, residential sales and business activities to show that this quasi experimental variation reflected in inequality of house prices and land use, influencing local development and urban sprawl. An instrumental variables strategy yields significant house price gains for increased perception of school quality, and lower deprivation in school neighborhoods. Our approach ensures improved external validity with respect to boundary discontinuity strategies.

11:30
12:00

The Long-Term Effects of Start-Up Subsidies for the Unemployed: New Evidence from Germany

Stefan Tubbicke - University of Potsdam

The German Start-Up Subsidy program for the unemployed recently underwent a major make-over, altering its institutional set-up, changing the composition of participants, and reducing overall participation rates. With its key parameters altered, ex-ante predictions on the program's effectiveness are ambiguous. Using propensity score matching, we provide estimates of long-term effects of the post-reform subsidy on individual employment prospects and labor market earnings up to 40 months after entering the program. Our results suggest large and persistent long-term effects of the subsidy on employment probabilities and net earned income. Extensive sensitivity checks with respect to the implementation of matching and the role of unobservable confounders reveal that our results are robust both to choice of algorithms used to create balanced samples in observable characteristics as well as deviations from the unconfoundedness assumption. The latter is tested using various approaches, including an instrumental variable strategy exploiting regional variation in the likelihood of receiving treatment.

12:00
12:30

One-off subsidies and long-run adoption - Experimental evidence on improved cooking stoves in Senegal

Gunther Bensch - RWI Leibniz Institute for Economic Research

Free technology distribution can be an effective development policy instrument if adoption is socially inefficient and hampered by affordability constraints. For improved cookstoves, this paper studies the effect of one-time free distribution on the willingness to pay in the long run. Using a randomized controlled trial, we find that households who received a free stove in the past reveal a higher willingness to pay to repurchase the stove. Learning effects thus at least compensate for potential reference-dependence effects. Our findings suggest that one-time free distribution does not disturb but might even rather facilitate future market establishment.

12:30-14.00 Lunch

14.00-15:30 SECOND SESSION: Unconfoundedness - New Methods and Applications

14:00
14:30

Subgroup Balancing Propensity Score

Junni Zhang - Peking University

We investigate the estimation of subgroup treatment effects with observational data. Existing propensity score matching and weighting methods are mostly developed for estimating overall treatment effect. Although the true propensity score should balance covariates for the subgroup populations, the estimated propensity score may not balance covariates for the subgroup samples. We propose the subgroup balancing propensity score (SBPS) method, which selects, for each subgroup, to use either the overall sample or the subgroup sample to estimate propensity scores for units within that subgroup, in order to optimize a criterion accounting for a set of covariate-balancing conditions for both the overall sample and the subgroup samples. We develop a stochastic search algorithm for the estimation of SBP-S when the number of subgroups is large. We demonstrate through simulations that the SBPS can improve the performance of propensity score matching in estimating subgroup treatment effects. We then apply the SBPS method to data from the Italy Survey of Household Income and Wealth (SHIW) to estimate the treatment effects of having debit card on household consumption for different income groups.

14:30
15:00

Smile: a Simple Diagnostic for Selection on Observables

David Slichter - Binghamton University

This paper develops a simple visual diagnostic for the selection on observables assumption in the case of a binary treatment variable. Under common assumptions, differences between treated and untreated observations in the unobservable determinants of treatment become very large among observations with propensity score close to 0 or 1. Therefore, if unobservables enter into the outcome equation, the estimated treatment effect will grow large for observations with propensity score close to 0 or 1. Researchers can check for this pattern by looking for a “smile” shape in a simple binned scatterplot. In empirical examples, the pattern is present when expected, and not present when not.

15:00
15:30

Wage Subsidies for the Unemployed: Does their Long-Run Effectiveness Change over Time?

Marina Furdas - Humboldt University

Employer-side wage subsidies are often regarded as a powerful short-run labor market instrument in stimulating hiring and facilitating reemployment among vulnerable groups of individuals. However, empirical evidence on their long-run effectiveness and on the question whether their future impact is related to the economic conditions at program start is rather scarce. This paper investigates the long-run effectiveness of a German labor market policy that combines employer-side wage subsidies with training on the future labor market prospects of unemployed individuals over time. In the empirical analysis we estimate calendar time-specific effects in a dynamic setting using two comparison groups for program participants. Our findings imply a reasonable time-specific variation in the long-run program effectiveness, which to some extent is decreasing over the years. Individuals receiving on-the-job training through the subsidy from 1993 onwards do not benefit from the program compared to similar unsubsidized workers. Further, we find only weak evidence for a positive relationship between program effectiveness and the economic conditions prevailing at program start. Finally, the program proves to be quite ineffective for participants with relatively short unemployment experience, thus suggesting potential for deadweight losses when the subsidy is granted to firms hiring newly unemployed individuals.

15:30-16:00 Coffee Break

16:00-17:00 THIRD SESSION: Theoretical Developments in Program Evaluation

16:00
16:30

Synthetic Control Estimation Method: A Generalized Inference Procedure and Confidence Sets

Sergio Firpo - INSPER

The Synthetic Control Estimation Method (SCEM) was proposed to answer questions involving counterfactuals when only one treated unit and a few control units are observed. SCEM has been widely used, despite the fact that its inference procedure has not yet been completely established. We contribute to inference in several fronts. We state sufficient assumptions that guarantee the adequacy of Fisher's exact hypothesis testing procedure for panel data. This procedure allows us to test any sharp null hypothesis and, consequently, to propose a new way to estimate confidence sets by inverting the test statistics. Moreover, we analyze the size and the power of the proposed tests with a Monte Carlo experiment and find that test statistics based on SCEM outperforms test statistics commonly used in the evaluation literature. We also extend our framework for the cases (i) when we observe more than one outcome of interest, (ii) more than one treated unit and (iii) when heteroskedasticity is present. Furthermore, we propose a sensitivity analysis that allows the researcher to verify the robustness of the empirical conclusions to some of our inference procedure's underlying assumptions. Finally, we apply our theoretical developments to reevaluate the economic impact of ETA's terrorism on the Basque Country.

16:30
17:00

Estimation and Inference of Treatment Effects with L_2Boosting in High-Dimensional Settings

Martin Spindler - University of Hamburg

Boosting algorithms are very popular in Machine Learning and have proven very useful for prediction and variable selection. Nevertheless in many applications the researcher is interested in inference on treatment effects or policy variables in a high-dimensional setting. Empirical researcher are more and more faced with rich data sets containing very many controls or instrumental variables where variable selection is challenging. In this paper we give results for valid inference of a treatment effect after selecting amongst very many control variables and instrumental variables estimation with potentially very many instruments when post- or orthogonal L_2Boosting is used for variable selection. We give simulation results for the proposed methods and an empirical application.

17:00
17:30

Copula-Based Random Effects Models for Clustered Data

Santiago Pereda Fernandez - Bank of Italy

In a binary choice panel data framework, when the unobserved heterogeneity is correlated across individuals, joint and conditional events depend on this correlation. In this setup, standard discrete choice panel data estimators do not provide consistent estimators of the probability of these events. I propose a random effects estimator that models the dependence among the unobserved heterogeneity of individuals in the same cluster using a parametric copula. This estimator allows to compute joint and conditional probabilities of the outcome variable, and I describe its properties, establishing its efficiency relative to standard random effects estimators, and propose a specification test for the copula. The implementation of the estimator requires the numerical approximation of high-dimensional integrals. To overcome the curse of dimensionality from which methods like Monte Carlo integration suffer, I propose an algorithm that works for Archimedean copulas. I illustrate this approach with an application of labor supply in married couples.

Friday, 14th July 2017

08:45-09:30 Welcome Coffee

09:30
10:30

From Local to Global: A Case Study in External Validity

Keynote speech by Rajeev Dehejia - New York University

We study issues related to external validity for treatment effects using 166 replications of the Angrist and Evans (1998) natural experiment on the effects of sibling sex composition on fertility and labor supply. The replications are based on census data from around the world going back to 1960. We decompose sources of error in predicting treatment effects in external contexts in terms of macro and micro sources of variation. In our empirical setting, we find that macro covariates dominate over micro covariates for reducing errors in predicting treatments, an issue that past studies of external validity have been unable to evaluate. We develop methods for two applications to evidencebased decision-making, including determining where to run the next experiment and whether policy-makers should commission new research or rely on an existing evidence base for making a policy decision.

10:30-11:00 Coffee Break

11:00-12:30 FOURTH SESSION: Spillovers and Interference in Program Evaluation

11:00
11:30

Encouragement, experience and spillover effects in a field experiment on teens’ museum attendance

Marco Mariani - IRPET

This paper revisits results from a field experiment conducted in Florence, Italy, to study the effects of incentives offered to high school teens to motivate them to visit art museums and to identify best practices to transform this behavior into a long run cultural consumption. Students belonging to a first group of classes receive a flier with basic information and opening hours of a main museum in Florence, Palazzo Vecchio. Students in a second group of classes receive the flyer and a short presentation conducted by an art expert. Students in a third group of classes, in addition to the flyer and the presentation, receive also a nonfinancial reward in the form of extra-credit points towards their school grade. Taking a Principal Stratification approach, we explore the causal pathways that may lead students to increase their future museum attendance. Within the strata defined by compliance to the three forms of encouragement, we estimate associative and dissociative principal causal effects, that is, effects of the encouragement on the primary outcome, long run cultural consumption, that are associative or dissociative with respect to the effects of the encouragements on the Palazzo Vecchio visit. This analysis allows to interpret these effects as ascribable either to the encouragements, or to the museum visit, or to classroom spillovers. To face identification issues, estimation is performed with Bayesian inferential methods using hierarchical models to account for clustering. The main findings of the analysis are as follows: what seems to matter the most is the motivational incentive (i.e., the presentation), rather than the induced experience, i.e., the Palazzo Vecchio visit. These results suggest that policies designed to change determinants of behaviour are more effective than policies designed to change behavior itself.

11:30
12:00

A framework for separating individual treatment effects from spillover, interaction, and general equilibrium effects

Andreas Steinmayr - University of Munich

This paper suggests a causal framework for disentangling individual level treatment effects and interference effects, i.e., general equilibrium, spillover, or interaction effects related to treatment distribution. Thus, the framework allows for a relaxation of the Stable Unit Treatment Value Assumption (SUTVA), which assumes away any form of treatment-dependent interference between study participants. Instead, we permit interference effects within aggregate units, for example, regions or local labor markets, but need to rule out interference effects between these aggregate units. Borrowing notation from the causal mediation literature, we define a range of policy-relevant effects and formally discuss identification based on randomization, selection on observables, and difference-in-differences. We also present an application to a policy intervention extending unemployment benefit durations in selected regions of Austria that arguably affected ineligibles in treated regions through general equilibrium effects in local labor markets.

12:00
12:30

Treatment Effects with Heterogeneous Externalities

Eleonora Patacchini - Cornell University

This paper proposes a new method for estimating heterogeneous externalities in policy analysis when social interactions take the linear-in-means form. We establish that the parameters of interest can be identified using specific functions of the share of the eligible population. We also show that the parameters can be consistently estimated, and we study the finite sample performance of the proposed estimators using Monte Carlo simulations. The method is illustrated using data on the PROGRESA program. We find that more than 50% of the effects of the program on schooling attendance are due to externalities, which are heterogeneous within and between poor and nonpoor households.

12:30-14.00 Lunch

14:00-15:30 FIFTH SESSION: Longitudinal Studies and Principal Stratification

14:00
14:30

Bayesian Inference in Longitudinal Regression Discontinuity Designs

Alessandra Mattei - University of Florence

We consider Regression Discontinuity (RD) designs where the treatment is dynamically assigned according to a sequence of cut-o_ rules based on a time-varying forcing variable. We use a probabilistic formulation of the assignment mechanism underlying RD designs making a local overlap assumption which accounts for the presence of the longitudinal forcing variable. The local overlap assumption defines subpopulations of units with a value of the time-varying forcing variable falling around the corresponding cut-o_ point. For these subpopulations we invoke a local latent sequential ignorability assumption to identify and estimate local causal effects of sequences of treatments. We propose a Bayesian approach to select the subpopulations and to draw inference on the target causal estimands. We apply our framework to a longitudinal study on the evaluation of Italian University student-aid policies on academic careers.

14:30
15:00

Bounding Average and Quantile Effects of Training on Employment and Unemployment Durations under Selection, Censoring, and Noncompliance

German Blanco - Illinois State University

Using data from a randomized evaluation of the Job Corps (JC) training program, we estimate nonparametric bounds for average and quantile treatment effects of training on employment and unemployment duration. Under relatively weak assumptions, we bound these effects addressing three pervasive problems in randomized evaluations: sample selection, censoring, and noncompliance. The first arises when the individuals' decision to experience employment or unemployment spells is endogenous and potentially affected by the program. Censoring arises when the duration outcome is fully observed only for individuals who have completed a full spell by the end of the observation period, with the extent of censoring being potentially affected by training. Noncompliance is present when some assigned participants do not receive training and some assigned nonparticipants receive training. Ignoring these issues would yield biased estimates of the effects. Our results indicate that JC training increases the average duration in weeks of the last complete employment spell before week 208 after randomization by at least 10.7 log points (11.3 percent) for individuals who comply with their treatment assignment and who would experience a complete employment spell whether or not they enrolled in JC. The proposed approach allows us to also bound the wage effects of JC for these individuals during those spells. We find that JC increases their average wages by between 6.2 and 13.7 log points (6.4 and 14.7 percent), suggesting that JC not only helps these individuals to maintain their jobs longer, but also that those jobs are better paid. We find no distinguishable effects of JC on average unemployment duration. The quantile results reflect heterogeneous effects and strengthen our conclusions based on the average effects.

15:00
15:30

Assessing Causal Effects in a longitudinal observational study with "truncated" outcomes due to unemployment and nonignorable missing data

Michela Bia - LISER

In this paper we analyze the short- and long-run effect of foreign language training programs on employment and wages measured over time, using administrative data on labour force in Luxembourg (IGSS-ADEM dataset). We develop a novel framework to simultaneously handle truncated wages due to unemployment, with incomplete observations not ignorable over time. In our study we find that language training programs increased re-employment probabilities, with no effect on the wages. This might be an incentive for the Employment Agency to better design future policies implemented in the context of language trainings. We then focus the analysis on the group of defiant-employees and find that defiers at 18 months switch to the always-employees stratum at 36 months with a proportion of almost 50% (the highest transition probability between the two periods). This evidence is in line with the economic theory: defiant-employees are subjects who accept any job, when not trained, but prefer to wait for a job with higher wage, when exposed to the program, because they feel better equipped.

15:30-16:00 Coffee Break

16:00-17:30 SIXTH SESSION: External Validity in Program Evaluationn

16:00
16:30

From Sample average treatment effect to population affect treatment effects on the treated

Richard Grieve - London School of Hygiene and Tropical Medicine

Randomized controlled trials (RCTs) can provide unbiased estimates of sample average treatment effects. However, a common concern is that RCTs may fail to provide unbiased estimates of population average treatment effects. We derive the assumptions that are required to identify population average treatment effects from RCTs. We provide placebo tests, which formally follow from the identifying assumptions and can assess whether they hold. We offer new research designs for estimating population effects that use non-randomized studies to adjust the RCT data. This approach is considered in a cost-effectiveness analysis of a clinical intervention: pulmonary artery catheterization.

16:30
17:00

Evaluating How Child Allowances and Daycare Subsidies Affect Fertility

Jian Li - University of Luxembourg

We compare the cost effectiveness of two pronatalist policies: (a) child allowances; and (b) daycare subsidies. We pay special attention to estimating how intended fertility (fertility before children are born) responds to these policies. We use two evaluation tools: (i) a dynamic model on fertility, labor supply, outsourced childcare time, parental time, asset accumulation and consumption; and (ii) randomized vignette-survey policy experiments. We implement both tools in the United States and Germany, .finding consistent evidence that daycare subsidies are more cost effective. Nevertheless, the required public expenditure to increase fertility to the replacement level might be viewed as prohibitively high.

17:00
17:30

End-of-Year Spending and the Long-Run Employment Effects of Training Programs for the Unemployed

Bernd Fitzenberger - Humboldt University

This study re-estimates the employment effects of training programs for the unemployed using exogenous variation in participation caused by budget rules in Germany in the 1980s and early 1990s, resulting in the infamous “end-of-year spending”. In addition to estimating complier effects with 2SLS, we implement a flexible control-function approach to obtain the average treatment effect on the treated (ATT). Our findings are: Participants who are only selected for budgetary reasons do not benefit from training programs. However, the ATT estimates suggest modest positive effects in the long run. Longer programs are more effective than shorter and more practice-oriented programs.

LISER

LUXEMBOURG INSTITUTE OF
SOCIO-ECONOMIC RESEARCH

www.liser.lu