Oxford Policy Management

Big Data, Machine Learning, and Causal Inference

ICT4Eval – International Conference
IFAD Headquarters, Rome, Italy

Paul Jasper
paul.jasper@opml.co.uk

6 June 2017

Contents

- Introduction
- What is new about Big Data?
- Statistical learning and machine learning approaches
- How can machine learning approaches be employed to help with causal inference?
- One example: preventing model misspecification in quasi-experimental evaluations
- Conclusion

Introduction

- Until very recently, Big Data and machine learning was not something most economists were concerned with

- But then, much attention was paid to Varian (2014) in the Journal of Economic Perspectives:

  “*In fact, my standard advice to graduate students these days is go to the computer science department and take a class in machine learning*.”

Big Data: New Tricks for Econometrics[†]

Hal R. Varian

Computers are now involved in many economic transactions and can capture data associated with these transactions, which can then be manipulated and analyzed. Conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools.

First, the sheer size of the data involved may require more powerful data manipulation tools. Second, we may have more potential predictors than appropriate for estimation, so we need to do some kind of variable selection. Third, large datasets may allow for more flexible relationships than simple linear models. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning, and so on may allow for more effective ways to model complex relationships.

In this essay, I will describe a few of these tools for manipulating and analyzing big data. I believe that these methods have a lot to offer and should be more widely known and used by economists. In fact, my standard advice to graduate students these days is go to the computer science department and take a class in machine learning. There have been very fruitful collaborations between computer scientists and statisticians in the last decade or so, and I expect collaborations between computer scientists and econometricians will also be productive in the future.

■ *Hal Varian is Chief Economist, Google Inc., Mountain View, California, and Emeritus Professor of Economics, University of California, Berkeley, California. His email address is hal@ischool.berkeley.edu.*

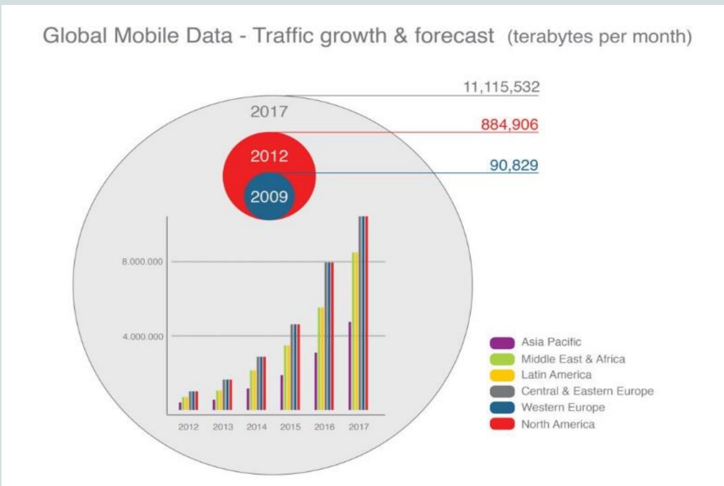
[†] To access the Appendix and disclosure statements, visit
http://dx.doi.org/10.1257/jep.28.2.3 doi=10.1257/jep.28.2.3

What is new about Big Data?

- Definition is unclear

…but it is certainly not just about the data.

New data?



Source: datapopalliance.org

'New' methods and tools!

Source: UN Global Pulse Projects

- Large N
- Large P
- Real time
- High frequency

- Computational statistics
- Artificial intelligence
- Machine/statistical learning
- 'Data Science'

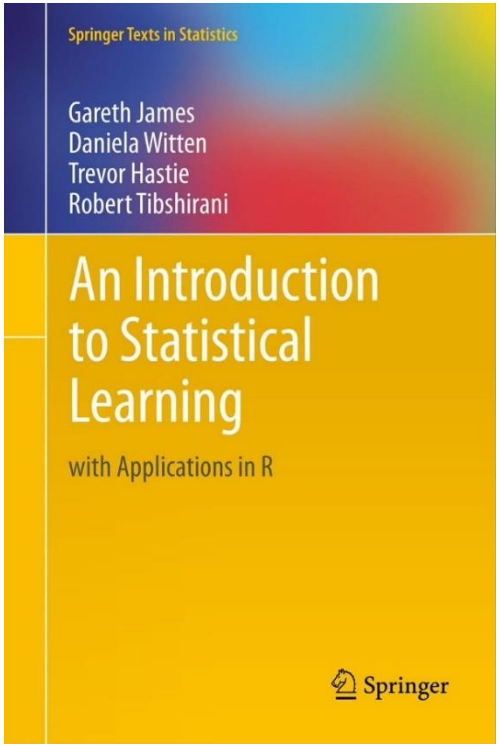
Statistical learning/machine learning vs. 'classical econometrics'

- "Understanding data"

- Since the late 1980s, "… statistical learning has emerged as a new subfield in statistics, focussed on supervised and unsupervised modelling and **prediction**".

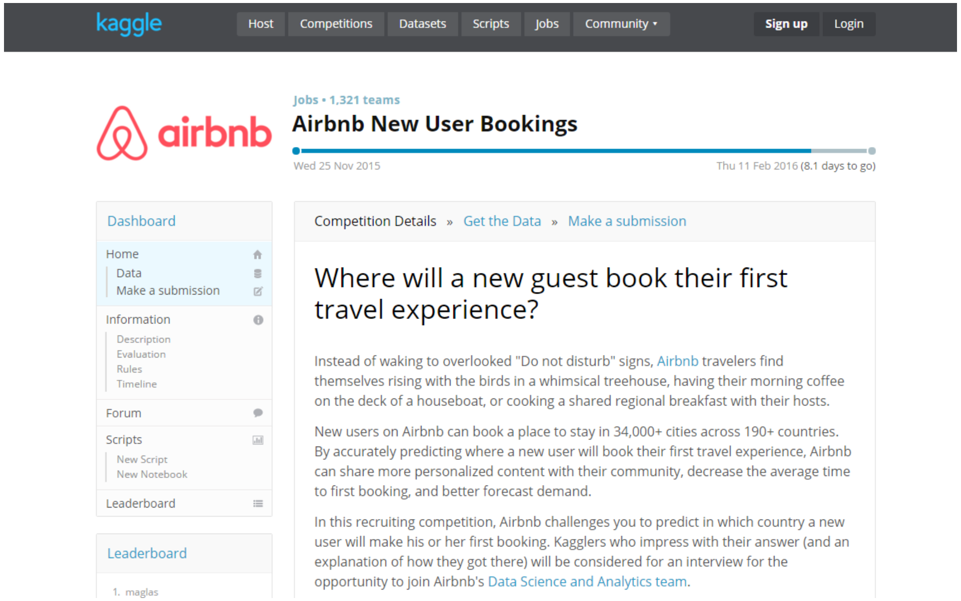
- An important distinction:
  - 'classical' methods in econometrics solve estimation problems
    - with assumptions about (linear) data generation processes (e.g. $Y = X\beta + \epsilon$) and distributions of variables involved
    - deriving algebraic solutions to estimation problems (e.g. OLS → $\hat{\beta} = (X'X)^{-1}X'Y$)
    - employing a frequentist approach to hypothesis testing
  - vs. computational methods that solve estimation problems
    - by exploiting **computational power in combination with re-sampling methods**
    - **derive highly non-linear, algorithmic solutions**
    - **and make little assumption about the data generating process**

→ I would consider methods that fall under the second definition as 'new' statistical learning in the narrow sense



Statistical learning: taxonomy of estimation problems

- Supervised learning:
  - Learn something about the relationship between features (x) and outcome measures (y), often **out-of-sample <u>prediction</u>** ($E(Y) = f(x)$) and **regularisation** (What are good predictors of y?)
- Unsupervised learning:
  - Spot patterns and structure in the data (only x data), often **summarisation**



Source: kaggle.com

Supervised learning approaches: some examples

- Regression trees:
  - Applicable for **non-linear prediction problems**
  - Re-sampling used to identify ‘depth’ of regression trees
- LASSO regression:
  - Quite well-known ‘penalised’ regression, where RSS is minimised subject to absolute value of estimated coefficients
  - Re-sampling used to identify ideal penalty term
  - Applicable for **regularisation, i.e. selection of best sub-set of explanatory variables**
- Vectors support machines
- Combinations!
  - Generally perform better than any singular predictor

→ **Re-sampling is always crucial to ‘tune’ models.**
  - Note that this requires computational power

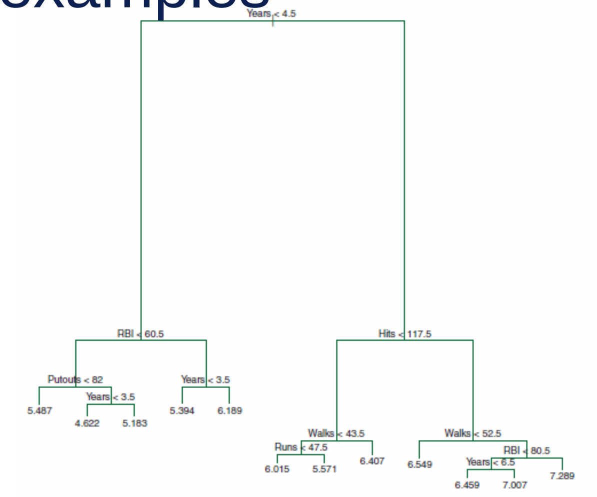
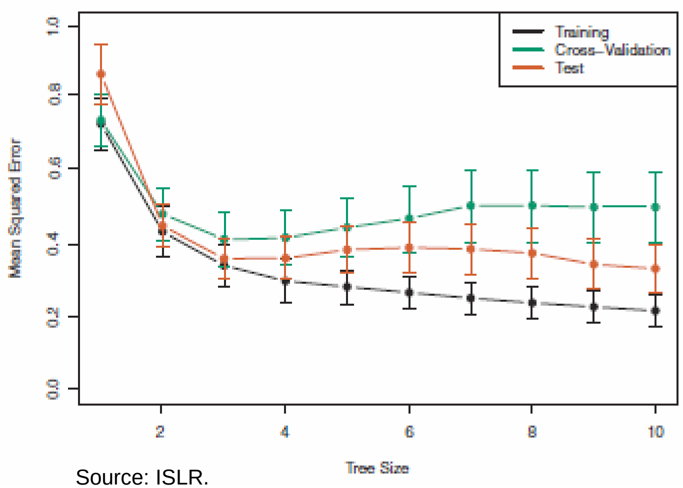


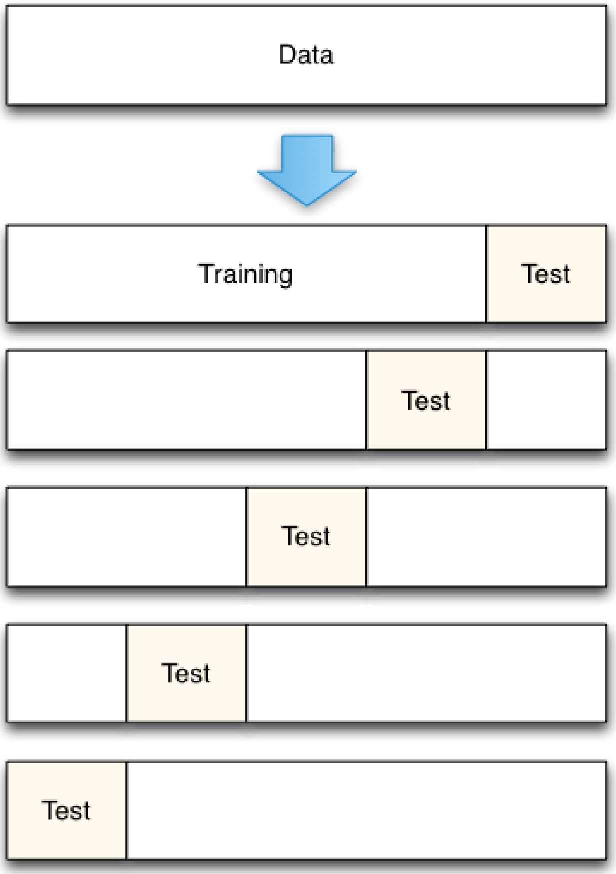
FIGURE 8.4. *Regression tree analysis for the* `Hitters` *data. The unpruned tree that results from top-down greedy splitting on the training data is shown.*



Source: ISLR.

Statistical learning: the importance of cross-validation for model fitting

- The basic idea **in a prediction context:**
  - Use part of your data to ‘**train**’ your model
  - Use the other part to ‘**test**’ it – *compare prediction to truth*
  - Repeat several times with different splits
  - Estimate your average performance (e.g. mean squared error)
- This can be employed both for
  - Model assessment
  - Model selection
    - Choose model specification that minimises your estimated MSE
    - Process avoids overfitting (in-sample performance vs out-of-sample performance)
- **Note: this is an entirely empirical method of choosing your best model.**



Source: kaggle.com

Statistical learning: strengths …

- Strengths:
  - Out-of-sample prediction and regularisation
  - ‘Deep learning’ - computer just won against human in Go
  - Can make use of all the Big Data around
  - Extremely powerful with large datasets

Statistical learning: strengths … and weaknesses

- Inference
  - No direct interest in parameter estimation
    - E.g. from a regression tree prediction, it is not possible to directly get $\hat{\beta}$
  - No direct interest in underlying data generation structure
    - Different prediction functions might have similar performance
  - Prediction is not <u>causal inference</u>
    - Predicting outcomes well does not directly help with counterfactual problem
    - **But – this is what we want to solve in impact evaluations!**

Artificial intelligence

Computer says Go

Beating a Go champion with machine learning

Jan 30th 2016 | From the print edition

Timekeeper Like 1.6k Tweet

How can machine learning be employed to help with causal inference?

- The distinction ML as prediction/regularisation tool vs causal inference is not really as clear-cut.
- There is an emerging literature on this topic:
  - American National Academy of Sciences had a colloquium on “Drawing Causal Inference from Big Data” in 2015
  - Justin Grimmer, Stanford, 2014: “We are all Social Scientists Now”
  - **Sendhil Mullainathan and Jann Spiess, JEP Spring 2017: “Machine Learning: an Applied Econometric Approach”**

SYMPOSIUM

We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together

Justin Grimmer, *Stanford University*

Information is being produced and stored at an unprecedented rate. It might come from recording the public's daily life: people express their emotions on Facebook accounts, tweet opinions, call friends on cell phones, make statements on Weibo, post photographs on Instagram, and log locations with GPS on phones. Other information comes from aggregating media. News outlets disseminate news stories through online sources, and blogs and websites post content and receive comments from their readers. Politicians and political elites contribute their own messages to the public with advertising during campaigns. The federal government disseminates information about where it spends money, and local governments aggregate information about how they serve their citizens.

The promise of the “big data” revolution is that in these data are the answers to fundamental questions of businesses, governments, and social sciences. Many of the most boisterous claims come from computational fields, which have little experience with the difficulty of social scientific inquiry. As social scientists, we may reassure ourselves that we know better. Our extensive experience with observational data means that we know that large datasets alone are insufficient for solving the most pressing of society's problems. We even may have taught courses on how selection, measurement error, and other sources of bias should make us skeptical of a wide range of problems.

This statement is true; “big data” alone is insufficient for solving society's most pressing problems—but it certainly can help. This paper argues that big data provides the opportunity to learn about quantities that were infeasible only a few years ago. The opportunity for descriptive inference creates the chance for political scientists to ask causal questions and create new theories that previously would have been impossible (Monroe et al. 2015). Furthermore, when paired with experiments or robust research designs, “big data” can provide data-driven answers to vexing questions. Moreover, combining the social scientific research designs makes the utility of large datasets even more potent.

The analysis of big data, then, is not only a matter of solving computational problems—even if those working on big data in industry primarily come from the natural sciences or computational fields. Rather, expertly analyzing big data also requires thoughtful measurement (Patty and Penn 2015), careful research design, and the creative deployment of statistical techniques. For the analysis of big data to truly yield answers to society's biggest problems, we must recognize that it is as much about social science as it is about computer science.

THE VITAL ROLE OF DESCRIPTION

Political scientists prioritize causal inference and theory building, often pejoratively dismissing measurement—inferences characterizing and measuring conditions as they are in the world—as “mere description” or “induction.” Gerring (2012) showed, for example, that 80% of articles published in *American Political Science Review* focus on causal inference. The dismissal of description is ironic because much of the empirical work of political scientists and theories that they construct are a direct product of description. Indeed, political scientists have developed a wide range of strategies for carefully measuring quantities of interest from data, validating those measures, and distributing them for subsequent articles. Therefore, although descriptive inference often is denigrated in political science, our field's expertise in measurement can make better and more useful causal inferences from big data.

The VoteView project is perhaps the best example of political science's expertise with measurement and why purely descriptive projects affect the theories we construct and the causal-inference questions we ask (McCarty, Poole, and Rosenthal 2006; Poole and Rosenthal 1997).[1] VoteView is best known for providing NOMINATE scores—that is, measures of where every representative to serve in the US House and Senate falls on an ideological spectrum. The authors are emphatic that NOMINATE measures only low-dimensional summaries of roll-call voting behavior. Like other measurement techniques, these summaries are a consequence of both the observed data and the assumptions used to make the summary (Clinton and Jackman 2009; Patty and Penn 2015). Extensive validations suggest, however, that the measures are capturing variation in legislators' expressed ideology (Clinton, Jackman, and Rivers 2004; Poole 1984; Poole and Rosenthal 1985; 1997).

The impact of the VoteView project is broad and substantial. NOMINATE measures appear in almost every paper about the US Congress and in much of the work of other scholars related to US politics. These findings have fueled numerous debates. Perhaps one of the most famous findings

80 PS • January 2015

doi:10.1017/S1049096514001784

How can machine learning be employed to help with causal inference?

Three main areas that I observe in evaluation work:

- Prediction as part of an estimation procedure: in many cases, causal inference requires a 'prediction step'
  - Instrumental Variables:
    - First stage: predict $\hat{x}$ using instruments (z)
    - Second stage: OLS on $y = \hat{x}\beta + \epsilon$
    - Hence, you can use ML to improve the first stage
  - Predicting counterfactuals when having 'big data'
    - E.g. UN Global pulse: "Using Financial Transactions Data to Measure Economic Resilience to Natural Disasters" (2016)

- Data-mining to find heterogeneous treatment effects and assess robustness of estimates to model selection
  - A lot of work by Susan Athey at Stanford

- **Regularisation to prevent model misspecification**
  - This is what we are looking at at OPM

The problem: model misspecification – an example

- Example taken from [Victor Chernozhukov's](#) **webpage (MIT)**
  - Note: 'big data' context with large p
  - Published in JEP, Spring 2014, Vol. 28 (2), pp. 29-50.

- Acemoglu, Johnson, Robinson (2001): *The effect of institutions on the Wealth of Nations*
  - Outcome: GDP per capita of countries today
  - Effect of interest: quality of institutions (D)
  - Instrument used: early settler mortality (Z)
  - Covariates:
    - Basic: constant and latitude
    - Flexible: transformations of latitude and continent dummies
- Problem: regularisation bias or OVB
  - Can we drop the flexible controls?

- Solution: double selection using Machine Learning (LASSO)
  - Select covariates that predict Y and
  - Select covariates that predict D or Z

→ This applies more generally to approaches that rely on controlling for observable covariates

	Institutions	
	Effect	Std. Err.
Basic Controls	**.96****	0.21
Flexible Controls	**.98**	0.80

Double Selection	**.78****	0.19

The problem: model misspecification in non-experimental impact evaluations

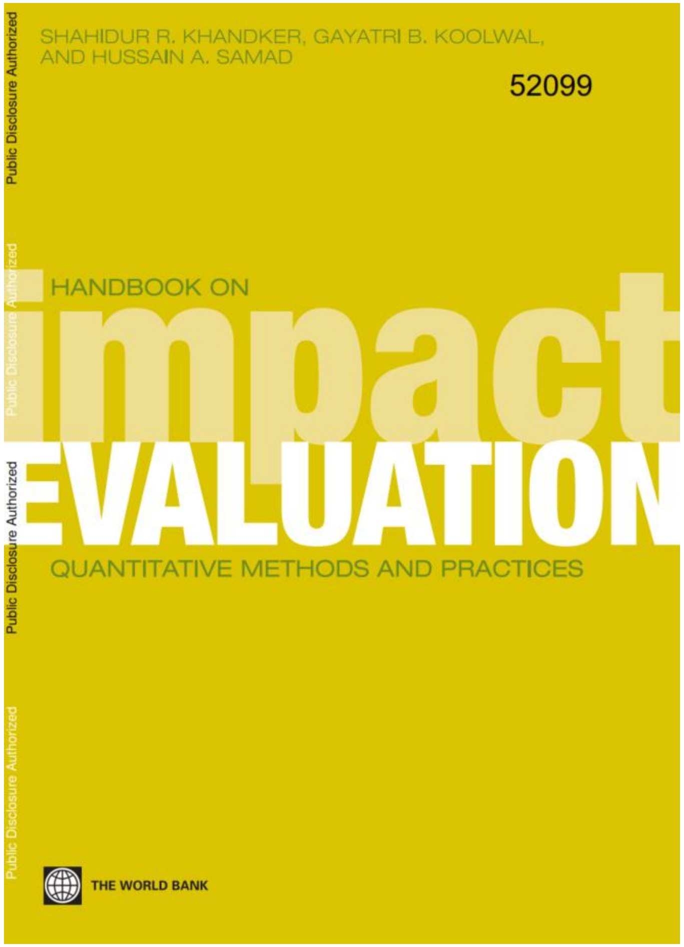
- In practice, many impact evaluations rely on conditional independence or unconfoundedness assumption:

  $$(Y_1, Y_0) \perp T | X$$

  - "Conditional on covariates, treatment assignment (T) is independent of the potential outcomes."

- This includes popular methods, such as e.g. PSM and regression approaches used in quasi-experimental evaluations

→ Selecting the right covariates, i.e. correct model specification is important to derive unbiased estimates of treatment effects

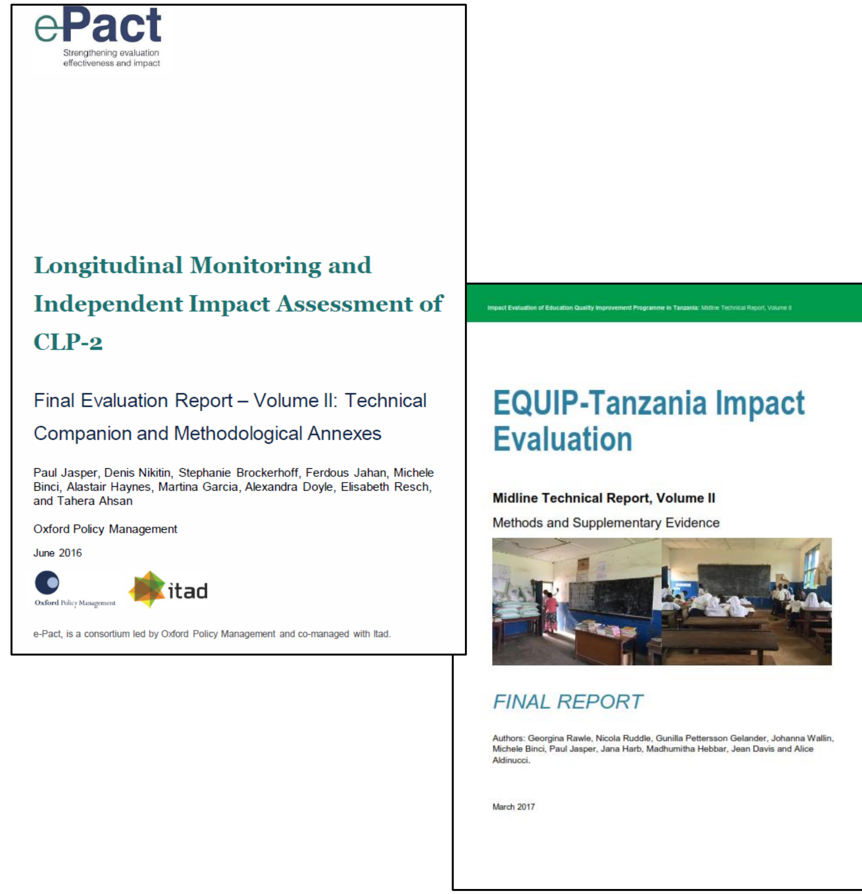


Preventing model misspecification in evaluations

- Approaches that are used for covariate selection:
  - Theory
  - ‘Deep knowledge’
- Potentially dangerous?
  - Survey data commonly has many variables (200+) – these can be combined in many ways to create ‘flexible controls’
  - It is likely that outcomes are related to covariates in complex, non-linear ways
  - Risk of omitted variable bias is large

- What is commonly done to address this:
  - Show robustness of results to different specifications

- What we are working on:
  - **Using algorithmic (stepwise regressions/LASSO) regularisation for principled model selection in quasi-experimental impact evaluations**
    - PSM approaches (EQUIP-T and CLP-2 Evaluations)



Preventing model misspecification: example of PSM

- PSM requires including the right set of covariates in the first-stage PS estimation
  - "Right set" means all covariates relevant to control for selection bias (i.e. related to treatment and outcome)
  - These can be non-linear transformations (polynomials, interactions) of basic covariates
  - Survey data often gives the possibility of controlling for 100+ covariates – together with transformations, this gives a very large set of potential covariates (large P).

- The approach we are testing (inspired by double ML literature):
  - Step 1: run algorithmic selection (stepwise regressions, LASSO) on both treatment and outcome, using basic covariates.
  - Step 2: repeat including transformations.
  - Step 3: predict PS using the union of selected variables.
  - Step 4: perform matching and balancing tests using different matching approaches


ePact
Strengthening evaluation effectiveness and impact

Longitudinal Monitoring and Independent Impact Assessment of CLP-2

Final Evaluation Report – Volume II: Technical Companion and Methodological Annexes

Paul Jasper, Denis Nikitin, Stephanie Brockerhoff, Ferdous Jahan, Michele Binci, Alastair Haynes, Martina Garcia, Alexandra Doyle, Elisabeth Resch, and Tahera Ahsan

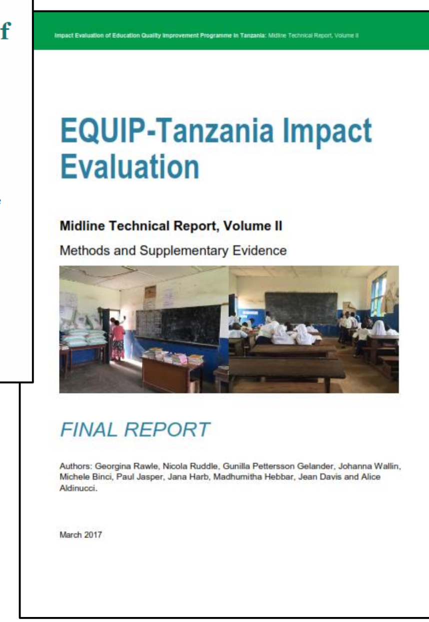
Oxford Policy Management

June 2016

Oxford Policy Management

itad

e-Pact, is a consortium led by Oxford Policy Management and co-managed with Itad.


EQUIP-Tanzania Impact Evaluation

Midline Technical Report, Volume II

Methods and Supplementary Evidence

FINAL REPORT

Authors: Georgina Rawle, Nicola Ruddle, Gunilla Pettersson Gelander, Johanna Wallin, Michele Binci, Paul Jasper, Jana Harb, Madhumitha Hebbar, Jean Davis and Alice Aldinucci.

March 2017

Conclusion: three main points

- Statistical/machine learning is here to stay - ‘classical’ econometric approaches will be mixed with computational methods more frequently for inference purposes.

- Much of this still sits in academic departments but slowly feeding into mainstream applied work.

- Some promising areas:
  - Predicting counterfactuals
  - IVs
  - In the context of RCTs: identifying heterogeneous treatment effects using principled data mining.
  - **Quasi-experimental and observational inference: preventing model misspecification.**

Source: https://memegenerator.net/instance/51894319

Oxford Policy Management

Thank you