Discussion: Can We Get More Out of Experiments?

Kosuke Imai

Princeton University

September 4, 2010

Keele, McConnaughy, and White

- Question: Can we gain efficiency by adjusting experimental data after the experiment is done?
- KMW's Answer: Yes, use matching rather than regression
  1. Much weaker functional-form assumption
  2. Can detect the lack of common support
  3. Less data snooping
- Disadvantages (acknowledged by KMW):
  1. May create imbalance in unobservables
  2. No design-based variance calculation
- KMW's proposal: report both unadjusted and adjusted estimates
- Adjust or not Adjust?: contribution to the important but controversial debate in the literature

Covariate Adjustments in Experiments

- Pre-randomization adjustments are gold standard
- Blocking *never* hurts (Imai, King & Stuart, 2008)
- Matching can hurt, but in practice it seems to work very well
- When post-randomization adjustments are desirable?
- Covariates are unavailable before randomization AND low power
  - Model-based variance calculation: this may be fine but not clear how to compare it with design-based variance
  - Risk of data snooping is always there
  - Which one do you trust if adjusted and unadjusted estimates are different?
- Some comments about details:
  1. Asymptotics: $\overline{T} \to 0$? maybe just refer to Freedman
  2. Simulation: Need to account for randomization?
  3. Randomization test: broken randomization?
  4. Empirical results: unadjusted $-0.00(0.822)$, with replacement $-1.25(0.039)$, without replacement $-0.25(0.803)$

Another Motivation for Covariate Adjustments

- Quantities of interest go beyond ATE
- Heterogenous treatment effects
  1. Useful for testing substantive theory
  2. Useful for policy-makers
- Growing methodological literature:
  1. Tree-based methods (Imai and Strauss)
  2. Generalized additive models (Feller and Holmes)
  3. Bayesian Additive Regression Trees (Green and Kern)
- Key challenge: avoid post-hoc subgroup analysis problem
- Regularization is required
  1. Cross-validation
  2. Bayesian prior
  3. Penalty function
- Using treatment effect heterogeneity to generalize experimental results to a larger population

Hartman, Grieve, and Sekhon

- Disadvantage of randomized experiments: external validity
- Question: How do we extrapolate from SATT to PATT?
- HGS's solution:
  1. Estimate heterogenous treatment effects via matching
  2. Weight pairs to match the population distribution
  3. Use placebo tests if possible
- Application to Pulmonary Artery Catheterization (PAC)
- Overall, a nice idea with an interesting application
- Some remaining issues:
  1. Variable selection problem: How should one choose variables to include in matching/weighting?
  2. Multiple testing problem with placebo tests
  3. Variance calculation is no longer randomization-based
- Suggestion: Use HGS's method with pre-randomization matching

Some Comments about Details

- Clarifying the identifying assumption:
  - Sample selection based on observables
  - Possibilities of unobserved confounders
- Bias decomposition:
  - Maybe helpful to decompose them into sample selection bias due to observables and unobservables
  - Should be expressed using potential outcomes, not $\mathbb{E}(Y_i \mid W, T_i = 1, I = 1)$ etc.
- Variance calculation:
  - Abadie & Imbens standard errors for SATE/SATT
  - What about PATT? Sometimes PATT has smaller standard error than SATT. Additional uncertainty due to sampling from population

Green and Kern

- Goal: Evaluate the performance of several competing estimators for generalizing SATE to PATE using Monte Carlo simulations
- Six methods
    1. Difference-in-means
    2. Linear regression with step-wise variable selection
    3. Inverse probability weighting (IPW)
    4. Genetic matching with maximum entropy weighting
    5. Bayesian Additive Regression Trees (BART)
- Use of realistic simulation settings based on GSS
- Linear, nonlinear response surfaces, confounded and unconfounded
- Findings:
    1. The difference-in-means is the worst
    2. BART often does better than the others
- Important contribution given the growing interest in the topic (Stuart et al.; Hartman et al.)

What Does Explain the Findings?

- No surprise that the diff-in-means performs badly
- No surprise that linear regression does badly
- Why does IPW do worse than BART?
  - IPW used here is parametric
  - Stabilized weights could be used
- Why does MaxEnt do worse then BART?
  - Common support assumption is satisfied
  - No variable selection for MaxEnt?

- Need for theoretical understanding about the conditions under which each model does and does not work well
- Report bias and efficiency rather than MSE

Back to the Common Theme

- Original question: Can we get more out of experiments?
- Yes, but be careful and use appropriate statistical tools

- Efficiency gain by pre-treatment covariate adjustments
- Post-treatment covariate adjustments require a greater care
  - Avoid post-hoc adjustment
  - Variable and model selection issues
  - Variance calculation

- Going beyond the SATE
- Heterogenous treatment effects and Extrapolation
  - Avoid post-hoc subgroup analysis problem
  - Variable and model selection
  - Sample selection based on unobservables

- Experiments vs. observational studies and central role of statistics
  - Internal vs. external validity
  - Small vs. large data sets