Experimental Methods in Social Sciences (in particular economics

Dirk Engelmann

Econometric Issues and Publication Bias

Exogenous Variation

- Key benefit of experiments is that variables of interest are varied exogenously
- Hence a lot of problems that econometrics addresses are solved
- But this is by far not the case for all interesting questions raised by experimental data
    - e.g., contributions and punishment decision in public good games are endogenous
    - standard econometric techniques are then applicable
- For some specific research questions structural estimates are useful
    - e.g., simultaneously estimating distribution of preference types and parameters for the different types (e.g., Cappelen et al., 2007)

Independent Observations

- Ideal case of experiment has many small independent units of observation
- This cannot always be achieved
  - e.g., repeated play with random matching typically turns whole session into single independent observation
    - matching in smaller matching groups (e.g., matching groups of 12 into groups of 4 in a session with 24 participants generates 2 independent observations per session)
    - but if we are really worried about effects of interactions, these should become worse in smaller matching groups
- Again standard econometric techniques are used to address this issue
  - panel data analysis
- Often researchers misuse corrections
  - clustering of standard errors with two few clusters (< 40)

Distributional Assumptions and Non-Parametric Tests

- In many cases applications of standard tests is problematic because experimental data is often far from normally distributed
  - e.g., in public good games large masses on 0 and full contributions
- Hence non-parametric tests are popular among experimentalists
- But these typically use aggregate data per independent observation because there is no correction for dependance
  - thus large number of independent observation is needed
- There is also wide-spread misunderstanding which hypothesis is tested by various tests
  - e.g., Mann-Whitney test is neither test of difference in means nor medians

Testing "Larger than 0"

- Experimental data often produces a specific problem
  - hypothesis predicts corner-point
    - e.g., price of 0, when negative prices are not possible or do not make sense
- This problem is rarely observed in standard empirical analysis
- Hence no standard econometric technique available
- Simply using, e.g., t-test makes no sense, because this assumes that errors are symmetrically distributed around corner-point prediction and hence rejects far too often
  - surprisingly many experimentalists do not know or understand this

Testing "Larger than 0"

- Solutions
  - Tobit regression without independent variables and check whether constant is significantly larger than 0
  - change design so that prediction is not corner-point anymore
  - sometimes one can test (and reject) alternative stronger hypothesis
    - e.g., when strategy space is $x \in [0, 100]$ test whether $x \leq 50$ instead $x = 0$
    - typically, this is a bit post-hoc
  - just stick with the observation that data is obviously larger than 0 (and not just 0 plus noise) without a formal test

Publication Bias and Statistical Power

- As all (empirical) science, experimental social science suffers from publication bias
  - significant (and interesting) results are more likely to be published
- In experimental social science, this problem may be made worse because many studies are underpowered (Maniadis, Tufano, and List, 2014)
  - for an individual study, still finding an effect even though it is underpowered is not such a bad thing because it suggests that the effect is strong

Publication Bias and Statistical Power

- For the universe of experiments, having typically underpowered studies implies that the share of false positives among all reported (significant) results increases
  - assume that half of expected (and tested) effects are real and the others are not
  - now assume we generally run tests with 5% significance level
  - hence 5% of the unreal effects will turn out to be significant
  - now assume that all tests have power q
  - then $q\%$ of all real effects turn out to be significant
  - then among all significant results, share of false positives is

$$\frac{\frac{0.05}{2}}{\frac{0.05}{2}+\frac{q}{2}}$$

  - clearly this is falling in q

Literature

- Cappelen, Alexander W., Hole, Astri D., Sørensen, Erik Ø, and Tungodden, Bertil (2007) "The Pluralism of Fairness Ideals: An Experimental Approach" American Economic Review 97(3), 818–827.
- Maniadis, Zacharias, Tufano, Fabio, and List, John A. (2014) "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects" American Economic Review 104(1), 277–290.