`DeclareDesign`: Simulation and Characteristics of Research Designs

Tara Slough

EGAP Learning Days IX: February 2018

Research Design Form and Research Design

- Research design form includes
  - Substantive details about your project (Section 1)
  - Elements of the research design
- Research designs exist independently of the application
  - A two-arm experiment with 50 units per arm randomized using simple random assignment (coin flip) could be used to study the effects of many treatments
  - Designs have statistical properties
  - We shoud assess a design by asking the question: "What we *could* we learn/have learned from the design?"

Four Elements of a Research Design

- Regardless of the method, research designs have four components
- MIDA:
  - *M*: Model (of how the world works)
  - *I*: Inquiry
  - *D*: Data strategy
  - *A*: Answer strategy

Model

- A model of how the world works, includes:
  - Endogenous variables
  - Exogenous variables
  - Functional relations between variables (potential outcomes)
  - Probability distribution over exogenous variables
- Somewhat more formal mapping of theory into the data than we are accustomed to
  - Can be thought of as a characterization of the data generating process (DGP)
- Note that this is fully general, not application-specific

Notes on the Model

- Where does it come from:
  - Theory
  - Past evidence
- The model is wrong by definition. If it were correct, you wouldn't need to do the study.
- But, without a model, we don't have a place to start in terms of assessing what *can* be learned

Inquiry

- An *answerable* question
  - Typically causal: What are the effects of Z on Y? What are the determinants of Y?
- Usually a quantity of interest, some summary of the data:
  - Descriptive: What is the mean of Y when some variable, $Z = 1$, formally $\mathbb{E}[Y|Z = 1]$?
  - Causal: What is the average difference of Y when $Z = 1$ versus when $Z = 0$, formally $\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]$?
  - Quantity is the *estimand*
- Not all questions that we want to ask are answerable
  - And the range of inquiries we can ask are limited: how much can we learn from some summary quantity?

Data

- Realize (generate) data on the set of variables
- A function of your *model*
- Includes both:
  - Sampling – how units arrive in your sample
  - Treatment assignment – what values of endogenous variables are revealed
- Fully general for all empirical social science:
  - Qualitative
  - Quantitative

Answer

- Given a realization of the data, generate an answer $\rightarrow$ an estimate of the quantity of interest (inquiry)
- This is your estimator:
  - Difference-in-means
  - Regression methods
  - etc.
- Answer is an estimate of the quantity of interest (inquiry/estimand)
  - We are used to looking at answers often without a clear inquiry

Logic of Simulation (Monte Carlo)

- With a model, inquiry, data (strategy), and answer we can simulate a research design.
- Mapping to standard Monte Carlo methods (roughly):
  - Model = Data generating process (DGP)
  - Inquiry = Estimand
  - Data = Realization of data generated by DGP
  - Analysis = Estimator
- As in standard Monte Carlo techniques:
  - Learn properties of estimator
  - Or, in this case, properties of research design

Properties of the Design

- Measured in terms of "diagnosands"
  - Well known: bias, RMSE, power
  - Plus many, many more...
- Designs can be compared or evaluated on the basis of statistical properties
  - Core insight: some research designs are better able to provide answers than others
  - For empirical work, we should use the best design available, given cost, ethical, and practical limitations

What does this mean for you?

- In parallel with the research design form, simulation will help you understand the properties of the design you are working on
  - A form of design "visualization"
  - Simulation allows for understanding of the consequences of different design choices
    - i.e. Do I need 300 instead of 200 units?
- `DeclareDesign` is a set of tools in R, including the packages:
  - `randomizr`: randomization, sampling etc.
  - `fabricatr`: generates data sets consistent with the specified model
  - `estimatr`: standard causal estimators, (computationally) efficient estimation

Design Form and `DeclareDesign`

- Design form: includes both substantive application and research design
  - Substantive content including question, motivation, context
  - Research design features, ideally refined by simulation
- Goal for the week:
  - Sufficient development of project such that you can fill in the research design form completely