Observational Study Designs I

POL-GA 3200
Quantitative Field Methods
Prof. Cyrus Samii
NYU Politics

September 21, 2016

Overview

Cochran (1983) defines an observational study as one in which,

1. *The objective is to study the causal effects of certain agents, procedures, treatments, or programs.*
2. *For one reason or another, the investigator cannot use controlled experimentation, that is, the investigator cannot impose on a subject or withhold from a subject, a procedure or treatment whose effects he desires to discover, or cannot assign subjects at random to different procedures.*

Overview

- No random assignment, so key concern is confounding.
- Observational study designs try to minimize confounding while also allowing for efficient estimation of causal effects.
- Common approaches for studying the effects of causes include matched designs and designs leveraging discontinuities or naturally “as-if random” assignment.
- Also designs attempting to identify causes of effects, known as “case-control” design or “choice-based” sampling design.

Overview

Part I:

- Designing a matched observational study.
- Designing a case-control study.

Matched Designs: Statistical Foundations

- Suppose a treatment, $D_i = 0, 1$ and potential outcomes, Y_{1i} and Y_{0i}, under treatment and control, respectively.

Matched Designs: Statistical Foundations

- Suppose a treatment, $D_i = 0, 1$ and potential outcomes, Y_{1i} and Y_{0i}, under treatment and control, respectively.
- There exists some natural process through which members of a population, $\mathscr{P}$, are assigned D_i values. Thus, by the time we arrive on the scene, the population has already been divided into those with $D_i = 1$ and $D_i = 0$.

Matched Designs: Statistical Foundations

- Suppose a treatment, $D_i = 0, 1$ and potential outcomes, Y_{1i} and Y_{0i}, under treatment and control, respectively.
- There exists some natural process through which members of a population, $\mathscr{P}$, are assigned D_i values. Thus, by the time we arrive on the scene, the population has already been divided into those with $D_i = 1$ and $D_i = 0$.
- Outcomes are eventually revealed, $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$.

Matched Designs: Statistical Foundations

- Suppose a treatment, $D_i = 0, 1$ and potential outcomes, Y_{1i} and Y_{0i}, under treatment and control, respectively.
- There exists some natural process through which members of a population, $\mathscr{P}$, are assigned D_i values. Thus, by the time we arrive on the scene, the population has already been divided into those with $D_i = 1$ and $D_i = 0$.
- Outcomes are eventually revealed, $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$.
- We set for ourselves the goal of estimating "average effect of the treatment on the treated" (ATT), defined as,

$$\rho_{ATT} = \mathrm{E}\left[Y_{1i} - Y_{0i} | D_i = 1\right]$$

  Everything is symmetric for the "average effect of the treatment on the controls" (ATC).

Matched Designs: Statistical Foundations

- Suppose a treatment, $D_i = 0, 1$ and potential outcomes, Y_{1i} and Y_{0i}, under treatment and control, respectively.
- There exists some natural process through which members of a population, $\mathscr{P}$, are assigned D_i values. Thus, by the time we arrive on the scene, the population has already been divided into those with $D_i = 1$ and $D_i = 0$.
- Outcomes are eventually revealed, $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$.
- We set for ourselves the goal of estimating "average effect of the treatment on the treated" (ATT), defined as,

  $$\rho_{ATT} = \mathrm{E}\left[Y_{1i} - Y_{0i} | D_i = 1\right]$$

  Everything is symmetric for the "average effect of the treatment on the controls" (ATC).
- Establishing the population for which you are making inferences is the crucial first step in a matched study design.

Matched Designs: Statistical Foundations

- The assumption necessary to identify ρ_{ATT} given that we can collect data from $\mathscr{P}$ is "conditional mean independence" (CMI) with respect to Y_{0i}:

  $$\mathrm{E}\left[Y_{0i}|D_i = 1, X_i\right] = \mathrm{E}\left[Y_{0i}|D_i = 0, X_i\right] \quad \text{and} \quad \Pr[D_i = 1|X_i] < 1,$$

  where X_i is a vector of covariates that we can measure on all members of $\mathscr{P}$, and $\Pr[D_i|X_i] < 1$ ensures that corresponding control units for all treatment units over the range of X_i.

Matched Designs: Statistical Foundations

- The assumption necessary to identify ρ_{ATT} given that we can collect data from $\mathscr{P}$ is “conditional mean independence” (CMI) with respect to Y_{0i}:

  $$\mathrm{E}[Y_{0i}|D_i = 1, X_i] = \mathrm{E}[Y_{0i}|D_i = 0, X_i] \quad \text{and} \quad \Pr[D_i = 1|X_i] < 1,$$

  where X_i is a vector of covariates that we can measure on all members of $\mathscr{P}$, and $\Pr[D_i|X_i] < 1$ ensures that corresponding control units for all treatment units over the range of X_i.
- Then,

  $$\begin{aligned} \mathrm{E}[Y_i|D_i = 1, X_i] - \mathrm{E}[Y_i|D_i = 0, X_i] &= \mathrm{E}[Y_{1i}|D_i = 1, X_i] - \mathrm{E}[Y_{0i}|D_i = 0, X_i] \\ &= \mathrm{E}[Y_{1i}|D_i = 1, X_i] - \mathrm{E}[Y_{0i}|D_i = 1, X_i] \\ &= \mathrm{E}[Y_{1i} - Y_{0i}|D_i = 1, X_i] \end{aligned}$$

  And,

  $$\int_x \mathrm{E}[Y_{1i} - Y_{0i}|D_i = 1, x] dF(x|D_i = 1) = \mathrm{E}[Y_{1i} - Y_{0i}|D_i = 1] = \rho_{ATT}.$$

Matched Designs: Statistical Foundations

- Go back to the expression for the ATT:

$$\rho_{ATT} = \mathrm{E}\left[Y_{1i} - Y_{0i} | D_i = 1\right] = \underbrace{\mathrm{E}\left[Y_{1i} | D_i = 1\right]}_{\text{observable}} - \underbrace{\mathrm{E}\left[Y_{0i} | D_i = 1\right]}_{\text{counterfactual}}$$

- Using CMI, we compute the counterfactual component by,

$$\mathrm{E}\left[Y_{0i} | D_i = 1\right] = \int_x \underbrace{\mathrm{E}\left[Y_{0i} | D_i = 0, x\right] dF\left(x | D_i = 1\right)}_{\text{observable}}.$$

Matched Designs: Statistical Foundations

- Furthermore, if we can come up with a matching solution, $\mathscr{M}$, consisting of weights to apply to the control units such that $F(x|D_i = 1) = F_{\mathscr{M}}(x|D_i = 0)$, we can compute the counterfactual component as,

$$\begin{aligned}\int_x \mathrm{E}\left[Y_{0i}|D_i = 0, x\right] dF(x|D_i = 1) &= \int_x \mathrm{E}\left[Y_{0i}|D_i = 0, x\right] dF_{\mathscr{M}}(x|D_i = 0)\\ &= \mathrm{E}_{\mathscr{M}}[Y_i|D_i = 0],\end{aligned}$$

  which is just the mean of the control units in data to which $\mathscr{M}$ is applied.

Matched Designs: Statistical Foundations

Examples of $\mathscr{M}$ satisfying the necessary conditions exactly:

- One-to-one exact matching: take each unit in the treated group, find a match on X_i in the control group, and allocate a weight of 1 to this matched control. Repeat for all treated units. (The weight values can accumulate.) For any control unit that is not matched, assign a weight of 0.
- Many-to-one exact matching: take each unit in the treated groups, find *all* K_i matches on X_i in the control group, and allocate a weight of $1/K_i$ to each of these matched controls. Repeat for all treated units. (The weight values can accumulate.) For any control unit that is not matched, assign a weight of 0.

Matched Designs: Statistical Foundations

- The solutions on the previous page assumed that exact matching was possible for all members of the treatment group.
- If this is not possible, approximations are done through:
  - "Nearest neighbor matching" using propensity scores, Mahalnobis distance, or some combination (e.g., GenMatch in the `MatchIt` or `Matching` packages).
  - "Coarsened exact matching" (CEM in the `MatchIt` package).
  - "Reweighting" that minimizes the discrepancy between $F(x|D_i = 1)$ and $F_{\mathscr{M}}(x|D_i = 0)$ over x (e.g., classical propensity score weighting, or `ebal` or `twang` packages).
  - A combination of the above.
- The further these approximations depart from the exact matching solution, the greater is the potential for bias.
- Estimation of causal effects proceeds as if we have a randomized experiment. So, covariate adjustment used to boost power.

Matched Designs: Implementation

So how do we turn these ideas into a *field* research design?

- If ATT is fixed as target estimand, basic research design is:
    1. Draw a representative sample from treated population, measure outcomes.
    2. Draw a matched sample from control population, measure outcomes.
- Even if we want to use reweighting, it is most efficient to draw control sample that is as matched as possible.
- Symmetric for ATC.

Matched Designs: Implementation

- Power analysis can be done by applying the same formulas that we used for designing an experiment.

Matched Designs: Implementation

- Power analysis can be done by applying the same formulas that we used for designing an experiment.
  - Typically we will stratify over combinations of X_i in sampling the treated and matched controls. Ignoring this in your power analysis would be conservative.

Matched Designs: Implementation

- Power analysis can be done by applying the same formulas that we used for designing an experiment.
  - Typically we will stratify over combinations of X_i in sampling the treated and matched controls. Ignoring this in your power analysis would be conservative.
  - Often matched observational studies examine effects of cluster-level treatments. Of course, the consequences of such clustering need to be taken into account.

Matched Designs: Implementation

- Power analysis can be done by applying the same formulas that we used for designing an experiment.
  - Typically we will stratify over combinations of X_i in sampling the treated and matched controls. Ignoring this in your power analysis would be conservative.
  - Often matched observational studies examine effects of cluster-level treatments. Of course, the consequences of such clustering need to be taken into account.
- Once you have determined your necessary sample size, you are ready to sample treated and then matched controls.

Matched Designs: Implementation

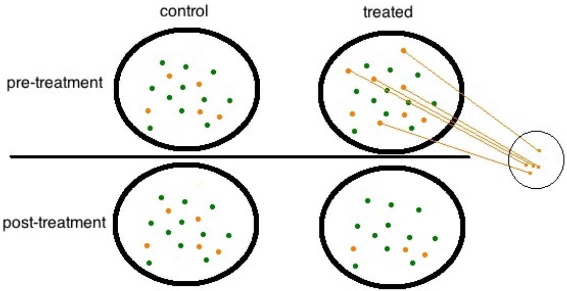
- Prior to matching you need covariate data that is sufficient for CMI to be believable. Be able to answer the question,

  For two units with the same value of X_i, how is it that they could differ in their D_i values? Is the reason something that is clearly innocuous in terms potential bias?

- Matching needs should be done on *pre-treatment covariates*.
  - Thus, data ideally comes from *pre-existing sources* that measured covariates prior to treatment assignment.
  - If the data are not available, you may proceed by collecting pre-treatment (and perhaps outcome) data on treated, then seek out matched controls.
  - Latter strategy has it's hazards:

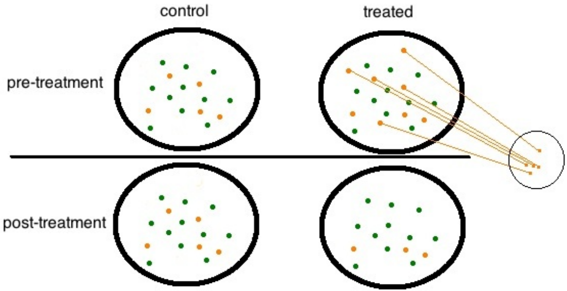
Matched Designs: Implementation

⚠Endogenous population composition change⚠:



Matched Designs: Implementation

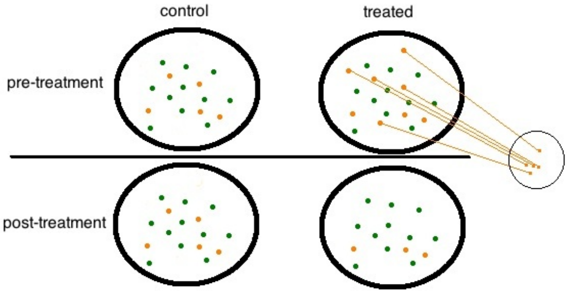
⚠Endogenous population composition change⚠:



- Post-treatment $D_i = 1$ population composition is not the same as pre-treatment $D_i = 1$ population composition, but latter is our target. This composition change is part of the *effect*.

Matched Designs: Implementation

⚠Endogenous population composition change⚠:



- Post-treatment $D_i = 1$ population composition is not the same as pre-treatment $D_i = 1$ population composition, but latter is our target. This composition change is part of the *effect*.
- With composition change effects, for your design to be unconfounded, X has to account for both treatment assignment and likelihood of relocating (akin to attrition problem in RCTs).

Matched Designs: Implementation

Suppose pre-existing data provide *some* of the information needed for CMI to hold, but not all. How to proceed? Suppose we want ATT.

Matched Designs: Implementation

Suppose pre-existing data provide *some* of the information needed for CMI to hold, but not all. How to proceed? Suppose we want ATT.

- First, clearly we want to match on what we can ahead of time.

Matched Designs: Implementation

Suppose pre-existing data provide *some* of the information needed for CMI to hold, but not all. How to proceed? Suppose we want ATT.

- First, clearly we want to match on what we can ahead of time.
- Second, it is intuitive that our design should have a control sample that is larger than our treated sample. But how much larger?

Matched Designs: Implementation

Suppose pre-existing data provide *some* of the information needed for CMI to hold, but not all. How to proceed? Suppose we want ATT.

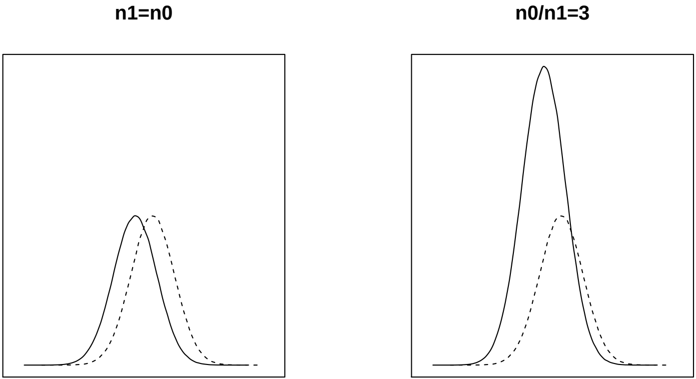
- First, clearly we want to match on what we can ahead of time.
- Second, it is intuitive that our design should have a control sample that is larger than our treated sample. But how much larger?
- One way to attack the problem formally is to think about how much you need to inflate the control sample to *ensure overlap on the margins* between the treated and control group on a given confounder. This will at least allow you to partially control for this confounder.

Matched Designs: Implementation

Suppose pre-existing data provide *some* of the information needed for CMI to hold, but not all. How to proceed? Suppose we want ATT.

- First, clearly we want to match on what we can ahead of time.
- Second, it is intuitive that our design should have a control sample that is larger than our treated sample. But how much larger?
- One way to attack the problem formally is to think about how much you need to inflate the control sample to *ensure overlap on the margins* between the treated and control group on a given confounder. This will at least allow you to partially control for this confounder.
- Consider the picture:

Matched Designs: Implementation



Matched Designs: Implementation

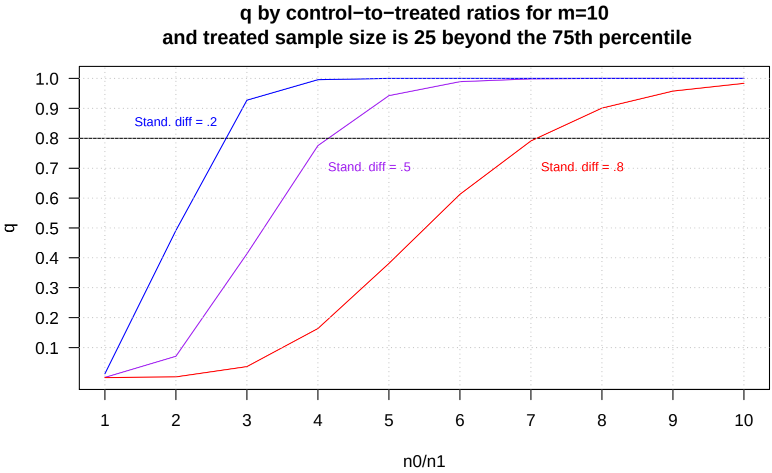
- This inflation can be expressed in terms of a ratio of control-to-treated sample sizes.
- Formally, suppose the confounder, C is distributed as $F(.;t,X)$, where t denotes treatment status and X any covariates used in stratifying our sampling. Consider a range of C values, $[c_L, c_H]$.
- The probability that a control unit value falls in this range is,

$$\pi \equiv F(c_H;0) - F(c_L;0)$$

- If we want our n_0 control units to have at least m observations in this range with probability q, then we choose the smallest n_0 such that,

$$q \leq \sum_{k=m}^{n_0} \binom{n_0}{k} \pi^k (1-\pi)^{n_0-k}.$$

Matched Designs: Implementation



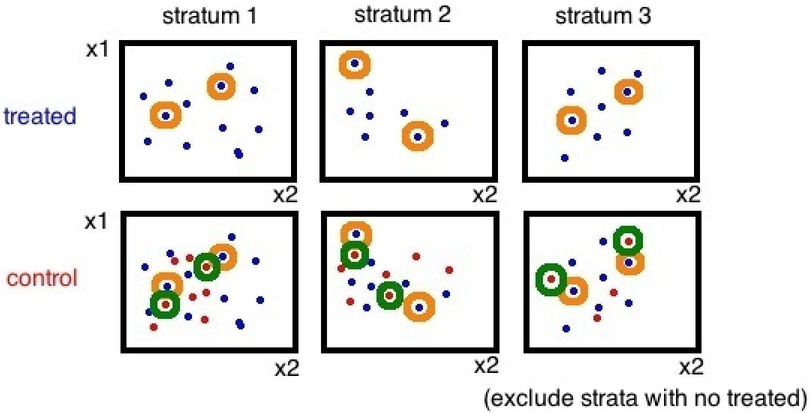
Example assumes *C* is standardized normal. The stand diff. measures the difference in treated and control means of the confounder, *C*.

Matched Designs: Implementation

Let’s look at two schematic examples:

Matched Designs: Implementation

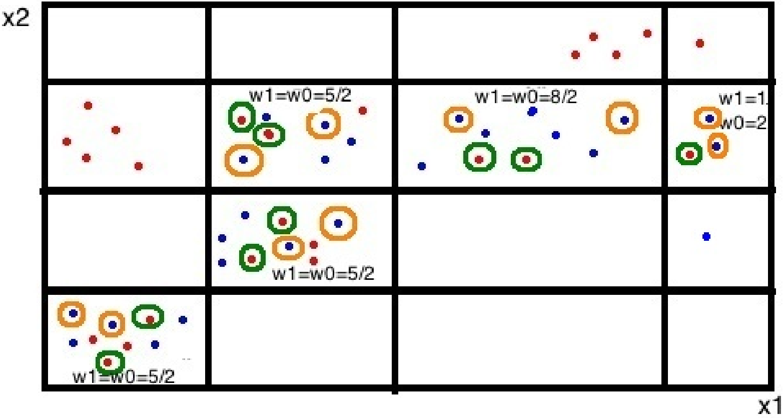
Example of nearest neighbor matching within strata:



For ATT, reweight within each stratum to recover the distribution of the *treated* population over the strata.

Matched Designs: Implementation

Example of a design using CEM:



ATT reweighting is shown for each stratification cell.

Case Control Studies

- A *case-control* study (also known as a *retrospective* study, *choice-based sample*, or *case-referent* study) flips the causal question around:

  What is the cause of an outcome that has already been revealed?
- Staple of epidemiology: what are possible causes of a disease outbreak?
- Useful when outcomes of interest occur rarely, take a long time, or were unintended.

Case Control Studies

- Formally, suppose a binary outcome, $Y_i = 0, 1$.
- A case-control study *selects on the dependent variable*, choosing a sample of units for which $Y_i = 1$ (“cases”) and another sample for which $Y_i = 0$ (“controls”).
- It compares cases and controls to determine the contributions of explanatory factors, $W_{1i}, ..., W_{Ki}$.
- If you focus on one explanatory factor, then you would proceed as with any other observational study, merely weighting to account for differential sampling rates conditional on outcome.

Case Control Studies

- Identification in such a study is delicate, because explanatory factors are usually embedded in endogenous causal chains that mask true causal relationships.
- Thus, case control studies rarely provide definitive answers. Rather, they provide leads that should be pursued via a more focused prospective experimental or quasi-experimental study.

Case Control Studies

Basic research design is:

1. Define a clear study population (e.g., age cohorts or demographic subgroups).
2. Obtain a representative sample of “cases” from the study population. This may be a sample stratified on factors thought to predict the outcome.
3. Obtain a sample of “controls” from the same study population. This may be done in a way that uses the same strata as the case sample, or you may select a control sample by *matching* to the cases.

Case Control Studies

- Matching in a case-control study typically increases efficiency/power to detect conditional effects.

Case Control Studies

- Matching in a case-control study typically increases efficiency/power to detect conditional effects.
- Matching followed by an unweighted analysis prevents you from studying causal effects of the matching variables (sample is balanced on these variables over the *outcome*).

Case Control Studies

- Matching in a case-control study typically increases efficiency/power to detect conditional effects.
- Matching followed by an unweighted analysis prevents you from studying causal effects of the matching variables (sample is balanced on these variables over the *outcome*).
- Matching following unweighted analysis prevents you from estimating the overall effect of an explanatory factor (effect heterogeneity needs to be aggregated in a way that accounts for the population distribution of the matching covariates).

Case Control Studies

- Matching in a case-control study typically increases efficiency/power to detect conditional effects.
- Matching followed by an unweighted analysis prevents you from studying causal effects of the matching variables (sample is balanced on these variables over the *outcome*).
- Matching following unweighted analysis prevents you from estimating the overall effect of an explanatory factor (effect heterogeneity needs to be aggregated in a way that accounts for the population distribution of the matching covariates).
- You can overcome these limitations by weighting to population distribution of matching covariates, accounting for differential sampling rates for cases and controls.

Case Control Studies

- Power analysis depends on the method of analysis.

Case Control Studies

- Power analysis depends on the method of analysis.
- Conventional presentations of case-control methods are based on logistic regression, owing to an invariance property on logistic slope coefficients. (See R code.) Then, power analysis would be based on the test that you use on the logit coefficients.

Case Control Studies

- Power analysis depends on the method of analysis.
- Conventional presentations of case-control methods are based on logistic regression, owing to an invariance property on logistic slope coefficients. (See R code.) Then, power analysis would be based on the test that you use on the logit coefficients.
- A more modern approach uses inverse propensity score weighting for *each risk factor of interest*, computes effects separately for each risk factor, and then uses multiple comparison adjustment to make a final judgment on the relative importance of difference risk factors (e.g., Van der Laan & Rose, Ch. 13-15; Young et al., 2009; Samii et al. 2014).