Daniels College of Business
UNIVERSITY OF DENVER

BEST PRACTICES FOR USING MATCHED SAMPLE DESIGNS TO REDUCE SELECTION EFFECTS

Donald Bergh; Louis D. Beaumont Chair of Business Administration and Professor of Management
Visiting Professor Erasmus University Rotterdam

Overview

- I will focus on selection bias (SB) and how sample matching can be used to most effectively reduce it prior to analytical remedies.
- Introduction: Definition, meaning, solutions, redirection, objective
- Overview of matched samples
- Alternatives for creating matched samples
- Best practice guidelines
- Draws from a working paper with my Erasmus coauthors Jose Gallegos Quezada, Rowen Moelijker, Ying Tang, and Max Welz

Introduction

- A major methodological challenge in organizational research is endogeneity
- Endogeneity is the correlation between *x* and *e* in a final sample. It is reflected in inconsistent estimate values.
- Different sources include omitted variables, measurement error, temporal relationships, **selection** among others
- Selection effects can “render coefficient estimates …uninterpretable” (Clougherty, Duso, & Muck, 2016: 287).

Introduction: Selection Definition

- Selection bias (SB) exists when we use samples rather than populations and the samples themselves may be unique from the population (e.g., non-random); it can also include self selection into a group (e.g., sample selection, self-selection)
- This bias is common in non-experimental research designs as they have no random assignment capability
- SB:
  - *Arises when a systematic difference exists between the sample and the population exists*
  - *Conducting research with a sample that does not accurately represent the population*

Introduction: Meaning

- SB occurs "when values of a study's dependent variable are missing a result of another process" (Certo et al., 2016: 2639); non-random samples
- Consider a study of stock market reactions to an acquisition announcement
  - *Stock market reactions are only available for firms that actually completed an acquisition, which become a nonrandom sample of firms*
  - *For firms not engaging in an acquisition, there are no corresponding stock market reactions to acquisition announcements*
- Firms making acquisitions may differ from those that do not, thereby representing a possible bias in the findings and how capital markets react to corporate event (selection effects)

Introduction: Solutions

- The common remedy:
  - *(1) Create a matched sample reflecting the larger population*
  - *(2) Pool the matched and focal samples together*
  - *(3) Empirically estimate whether an observation in the pooled sample (or population) appears in the focal sample (AKA 1st stage)*
  - *(4) Empirically estimate the hypothesized relationships including a term from the first test (AKA 2nd stage)*

Introduction: Solutions (cont'd)

- The literature now includes a widely accepted set of statistical resolutions
- Focusing on the use of instrumental variables and
- A system of estimating equations such as Heckman 2-stage and 3-stage least squares regression procedures

Introduction: Solution (cont'd)

- Analytical solutions focus on bias **after** the data have been collected; however, the research design is also relevant for SB
- The data collection structure is an "upstream" source of bias for the data set
- Redirect attention to the research design
  - *Research design may play a stronger role than statistics in addressing threats to the validity of findings*
  - *"…in the interplay between design and statistics, design rules" (Shadish, Cook, & Campbell, 2002: xvi).*

Introduction: Redirecting

- **We have very little consideration of creating, using, evaluating, and reporting matched sample studies in management research**
- It seems plausible that some studies have:
  - *Used matching procedures indiscriminately without justification*
  - *Produced matching samples that do not align closely with their study features*
  - *Dropped cases unnecessarily*
  - *Empirical findings and conclusions may have been adversely impacted*

Introduction: My Objective

- My Objective: A non-technical guide to using matched sample designs

- Recognize that research designs are a source of bias arising from sample- and self-selection
- Researchers benefit from knowing how to structure their designs to reduce bias before they attempt to correct for it in their statistical models
  - *Reduce bias in the research design*
  - *Less bias to correct in statistical analyses*

Overview of Matched Samples

- Non-experimental designs do not offer a random assignment opportunity;
  - *Sample consists of "treatment" behavior of interest, whether sampled on that treatment of self-selected into it.*

- Researchers can reduce this bias their creating control or comparison groups
  - *Creating these groups involves a process known as "matching", defined as "any method that aims to equate (or "balance") the distribution of covariates in the treated or control groups" (Stuart, 2010: 1).*
  - *The researcher "groups units with similar scores on the matching variable, so that treatment and control groups each contain units with the same characteristics as the matching variable" (Shadish et al. 2002: 118).*

Overview (cont'd)

- The primary reason for using matching is to reduce post-treatment differences in the outcome variable between treatment and control groups which might be attributable to the effect of the treatment or the pre-treatment differences between the groups (or a combination thereof).

- Matching strives to create a control sample that can decrease the effects of sample and treatment assignment selectivity, allowing researchers to consider their data as if it had come from a randomized experiment

Overview (cont'd)

- Matching reduces selection-based endogeneity, which can arise from two sources: sample selection and self-selection bias.
  - *Sample selection bias occurs when a sample is not representative of a true population*
  - *Self-selection bias arises when agents make choices regarding "assignments into mutually exclusive treatment and nontreatment groups based on unobservables that correlate with both outcomes and observable predictors" (Clougherty et al., 2016: 288).*

- Matching would create a comparison sample to estimate a direct effect of a certain treatment on a certain outcome variable. Many different methods exist for deriving matches, which are collectively referred to as "matching methods".

Overview (cont'd)

- Matching sample process encompasses four stages (Stuart, 2010):
- (1) defining "closeness" in terms of some conception of distance or differences;
- (2) implementing a matching method, given a chosen approach to operationalizing distance;
- (3) assessing the quality of the resulting matched samples, and if needed, repeating steps 1 and 2 until well-matched samples are produced; and
- (4) assessment of the outcome and estimation of the treatment effect, given the matching from step (3).

Overview (cont'd)

- Thus, the researcher identifies relevant covariates whereby a given treated study subject ("treated unit"') is paired with one or more control units ("nontreated units") that are as similar as possible (e.g., achieve "balance").

- This pair is a "match," and the matching algorithm attempts to determine as many matches of sufficiently high quality as possible.

- Matching strives to produce a similar distributional value of covariates between the treatment and control group, known as *covariate balance.*

Dominant Matched Sample Alternatives

- **Propensity score matching (e.g., distance measure)**
- PSM constructs a control group by matching each treated unit with a non-treated unit of similar attributes.
- **These matches allow the researcher to then estimate the impact of the treatment.**
- The PSM procedure strives to “reduce imbalance in pre-treatment covariates between the treated and control groups, so to reduce the degree of model dependence and bias” (King & Nielsen, 2019: 435).

Alternatives: Propensity Scores (cont'd)

- This method relies on the propensity score, which is a conditional probability of receiving a treatment, given the other observed explanatory variables.
- PSM utilizes a key finding reported by Rosenbaum and Rubin (1983) whereby adjusting for propensity scores eliminates selection bias stemming from treatment selectivity
- Propensity scores are unobserved and are estimated by regressing the treatment assignment variable T on the covariates X by means of linear logistic regression (Stuart, 2010).
- Binary classifiers that estimate individual class membership probabilities can be used for the estimation of propensity scores

Alternatives: Propensity Scores (cont'd)

- For each treated unit, the algorithm specifies the control unit with the closest estimated propensity score using a distance measure. The researcher will need to select one.

- Typically, a "nearest neighbor" algorithm attempts to find the *k* nearest neighbors of each unit for which to match. Careful attention to the choice of covariates to assess equivalence of distributional properties is needed to ensure sufficient stratification and balancing the propensity scores (Dehejia & Wahba, 1999).

- The more balance achieved between the two groups, the more efficient and less bias in the estimators. Matching on propensity scores can reduce case numbers.

Alternatives: Propensity Scores (cont'd)

- In general, PSM is a flexible matching method that has become popular across the social sciences
- However, the requirements and assumptions underlying PSM can limit its abilities to produce valid matches.
- King and Nielsen (2019) argue that PSM has multiple weaknesses including producing an inferior randomized experiment design to those of other matching models and may increase rather than decrease the degree of imbalance and incongruence among groups
- In those cases, other techniques such as coarsened exact matching become plausible

Alternatives: Coarsened Exact Matching (cont'd)

- CEM recodes the covariate values of two similar units to be exactly equal to one another: If a treated unit and a control unit have similar covariate realizations, CEM generates new coarsened covariates which are exactly equal for the two units.

- Iacus and colleagues (2012) explain: "The basic idea of CEM is to coarsen each variable by recoding so that substantively indistinguishable values are grouped and assigned the same numerical value (groups may be the same size or different sizes depending on the substance of the problem).

- Then, the "exact matching" algorithm is applied to the coarsened data to determine the matches and to prune unmatched units. Finally, the coarsened data are discarded and the original (uncoarsened) values of the matched data are retained" (page 8).

Alternatives: Coarsened Exact Matching (cont'd)

- After all covariates have been coarsened, the sample is partitioned into strata.
- All units within a given stratum are characterized by having the exact same values for all covariates. All treated units and control units within a stratum are considered matches.
- Based on these matches, specific CEM weights are used to calculate the treatment effect estimate.
- Coarsening works variable-by-variable, so researchers must balance coarsening each variable relative to losing information

Alternatives: Coarsened Exact Matching (cont'd)

- In some situations, measuring a variable in a coarsened way is pre-determined (e.g., CEO gender)
- In other situations, researchers measure certain variables in a coarsened way. For instance, personality traits are typically measured by means of Likert scales (e.g. agree, neutral, disagree). Categorical variables (ordered or non-ordered) typically do not require further coarsening
- Continuous variables are coarsened by partitioning the observed range of a covariate into multiple intervals of equal length.
- The degree of coarsening needs to be carefully chosen by the researcher and should be reported due to its importance in finding matches.

Alternatives: Coarsened Exact Matching (cont'd)

- CEM has been described as simple
- Computationally more efficient than alternative matched sample approaches
- Indeed, in head-to-head comparisons with other matched sample techniques, particularly PSM, CEM produced matched data sets with lower imbalance and larger sample sizes
- However, CEM does not allow for estimations of **average** treatment effects (ATE) and can lead to high bias and low precision due to reductions in sample sizes when larger numbers of covariates are used for matching

Alternatives: Other Considerations

- 1) Distinguish between matching with and without replacement.
- In matching with replacement, a control unit can be matched to more than one treated unit, whereas a control unit can only be matched to one single treated unit in matching without replacement.
- In matching without replacement, the order of the units in the sample affects the matches; hence the order of units should be randomized in this case.
- Matching with replacement generally leads to matches of higher quality (lower bias) at the cost of higher variance, while the reverse occurs in matching without replacement (e.g., smaller matched-sample with large variance)

Alternatives: Other Considerations (cont'd)

- 2) Both PSM and CEM require the assumption of ignorable treatment assignment ("*ignorability*") to hold true.
- This assumption, also known as *unconfoundedness*, imposes that *all* variables that *both* affect treatment assignment T and outcome Y have been measured and are available to the researcher in the covariate vector X
- In other words, ignorability requires that omitted variable bias is absent
- If we omit a variable that affects both the treatment and the outcome, we cannot expect to accurately capture the treatment's effect on the outcome.

Alternatives: Other Considerations (cont'd)

- 3) The ignorability assumption is not empirically testable and instead relies on a good case-specific understanding by the researcher and that all relevant covariates have been assessed.

- First develop theoretical understanding of the phenomenon or effect before resorting to methods for causal inference such as PSM or CEM.

- Such an understanding can be obtained by identifying and studying relevant and recent literature on the phenomenon of interest.

Alternatives: Other Considerations (cont'd)

- 4) A final assumption pertains to PSM.
- This assumption, called *overlap* or *positivity* assumption, requires that propensity scores are never equal to exactly 0 or 1;
- The probability of being assigned in the treatment group can never be 0 or 1 for any covariate level
- No observations are excluded from ever receiving treatment and that no individuals always receive treatment by design

Guidance

- Objective
- Covariate selection
- Choice of matching technique
- Assessing match quality
- Disclosure

Guidance (cont'd)

- **OBJECTIVE**
- Effect type to estimate and the population to which the effect generalizes.
- Matching uses a sample to estimate a direct effect of a certain treatment on a certain outcome variable.
- The estimated effect is expected to generalize to a certain target population, known as the population effect.
- Matching methods can estimate multiple population effects including the average treatment effect (ATE) and the average effect in the treated group (ATT).
- Standard software allows the researcher to *a priori* specify whether they wish to estimate an ATE or ATT.

Guidance (cont'd)

- **Covariate Selection**

- Ignorability assumption: "there are no unobserved differences between the treatment and control groups, conditional on the observed covariates…include in the matching procedure all variables known to be related to both treatment assignment and the outcome" (Stuart, 2011: 5).

- A covariate *should not* be affected **by the treatment of interest**, either directly or indirectly through another possibly unobserved variable. Selecting a covariate that has been influenced by the treatment intervention likely introduces substantial bias in any treatment effect estimate (Rosenbaum, 1984).

Guidance (cont'd)

- **Covariate Selection (cont'd)**

- Choose covariates that have been measured prior to the treatment intervention
- Conversely, a covariate may affect the treatment status.

Guidance (cont'd)

- Choosing a Matching Technique
- No single superior matching method in all cases

- **Concern over statistical bias**: CEM has less potential due to not involving distance calculations
- **Covariate imbalance:** CEM is less vulnerable to this problem
- **Number of covariates**: PSM accommodates more variables more effectively
- **Estimating ATE**: PSM is superior (CEM does not reflect population and is not applicable)

Guidance (cont'd)

- Assessing Match Quality
- Matching is deemed successful when it achieves covariate balance and many (most) treated units can be matched
- **Assess covariates across samples.**
- Compare covariate distributions between the treatment and control groups.
- Calculate the difference in the empirical means of each covariate between the treatment group and the covariate group, standardized by the pooled standard deviation across both groups (a.k.a. *standardized mean difference* (*SMD*)).
- Good balance is indicated by SMDs close to zero. Ho et al. (2011) recommend thresholds of 0.1.

Guidance (cont'd)

- **Assess the remaining sample size after matching**
- A trade-off between covariate balance and the remaining sample size
  - *Some treated units may not receive a match, and unmatched units are subsequently dropped from the analysis.*
  - *If there are relatively few matched units, the effect estimate may be imprecise, so it is preferable to have a larger post-matching sample size.*
- Achieving good balance may require the researcher to balance the region of common support of the two groups; narrower regions mean more cases are dropped, resulting in a smaller sample size.
- Re-do matching specifications in some instances

Guidance (cont'd)

- **Disclosure**
- Reporting covariates used in the matching process
- Documenting sample sizes in the initial and final matching process
- Describing matching method and its applicable algorithm
- Offering a justification for selecting a particular approach.

- Comparing the covariate values (e.g. means, standard deviations) across the matched samples will help establish the degree of balance among the groups and establish confidence that unique qualities of the treatment group have been reduced.

Summary

- Non-random sampling is a common issue in non-experimental designs
  - *What the sample consists of, how subjects got into a sample; create specter of 'sample selection bias'*
- Can attenuate correlation between independent variable(s) and error term
- Solutions include creating a matched sample and using a 2- or 3-stage analytical approach; most knowledge focuses on analytical decisions
- Reduce sample selection bias by increasing understanding of matching methodology and its decisions; also increase transparency of procedures
- In the interplay between design and analysis, "design rules" (Shadish et al 2022) – reduce bias in the creation of the data set

References

- Certo, S. T., Busenbark, J. R., Woo, H. S., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal*, *37*(13), 2639-2657.
- Clougherty, J. A., Duso, T., & Muck, J. (2016). Correcting for self-selection based endogeneity in management research: Review, recommendations and simulations. *Organizational Research Methods*, *19*(2), 286-347.
- Dehejia, R. H. and Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics*, 84, 151-161.
- Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1), 1-24.
- King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4), 435-454.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1-21.