SISCER Module 12
Lecture 2: Matching for Cohort Studies

Ting Ye & Qingyuan Zhao

University of Washington & University of Cambridge

July 2023

Plan

Key references for this lecture

- Holland (1986); Rosenbaum (2002, Chapter 1) for observational studies and causal inference
- Rosenbaum (2010, Chapters 8-13) for matching

What are observational studies? (Holland, 1986; Rosenbaum, 2002)

An observational study is an empiric investigation in which (Cochran, 1965)

> *. . . the objective is to elucidate cause-and-effect relationships . . . [in which] it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatment whose effects it is desired to discover, or to assign subjects at random to different procedures.*

Features of observational studies:

1. Study the effects of causes relative to other causes (Holland, 1986)
   - Not everything can be a cause; "No causation without manipulation"
2. Non-interventional: the investigator does not control the assignment of treatments

Example 1: smoking and lung cancer

- By the mid-1940s, it had been observed that lung cancer cases had tripled over the previous three decades. But the cause for the increase in lung cancer was unclear and not agreed upon.
- Possible explanations included
  - Changes in air quality due to introduction of the automobile
  - Widespread expansion of paved roads that contained many carcinogens
  - Aging of the population
  - The advent of radiography
  - Better clinical awareness of lung cancer and better diagnostic methods
  - Smoking

Example 1: smoking and lung cancer (cont.)

- A series of observational studies reported overwhelming association between smoking and lung cancer.
- Hammond (1964) matched 36,975 heavy smokers to nonsmokers on the basis of age, race, nativity, rural versus urban residence, occupational exposures to dust and fumes, religion, education, marital status, ...
- Of the 36,975 pairs, there were 122 discordant pairs in which exactly one person died of lung cancer. Among them,
  - 12 pairs in which nonsmoker died of lung cancer
  - 110 pairs in which smoker died of lung cancer

  So smoking is very strongly associated with lung cancer.
- Observational studies established the direct evidence linking smoking with human health.

We will study how to make inferences from this observational study.

Example 2: effect of vitamin C on cancer

- In a 1976 study, Pauling and Cameron presented observational data concerning the use of vitamin C as a treatment for advanced cancer.
- They gave vitamin C to 100 patients believed to be terminally ill from advanced cancer. For each such patient, 10 historical controls were selected of the same age and gender, the same site of primary cancer and the same histological tumor type.
- The vitamin C patients were reported to have a mean survival time about 10 months longer than that of the controls. Cameron and Pauling concluded "there is strong evidence that treatment. . . [with vitamin C]. . . increases the survival time."
- To test the claim, the Mayo clinic (Moertel et al., 1985) conducted a double blind RCT comparing vitamin C to placebo for patients with advanced cancer of the colon and the rectum. They found no indication that vitamin C prolonged survival.

The importance of reliable observational studies

Observational studies ...

- can provide evidence on critical questions that cannot be addressed by experiments
- can provide timely evidence at low cost in a real-world setting
- if not conducted with caution, can be dangerous in leading investigators to advocate harmful policies or ineffective treatments

Some general principles

- Design trumps analysis (Rubin, 2008)
  - Design: choose an identification strategy and collect relevant data
  - Analysis: apply an appropriate statistical method to analyze the data
- Randomization is the gold standard of causal inference, both for design and analysis.
  - "The planner of an observational study should always ask himself the question, How would the study be conducted if it were possible to do it by controlled experimentation?" (Dorn, 1953; Cochran, 1965)
- Quote from Fisher: "Make your theories elaborate."
  - "... when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold... this multi-phasic attack is one of the most potent weapons in observational studies." (Cochran, 1965)

Languages for causality

$$\text{Causal inference} \approx \text{Causal language/model} + \text{Statistical inference}$$

Three languages that are basically equivalent and advantageous for different purposes.

1. Potential outcomes/counterfactuals
2. Structural equations
3. Graphs

Randomized experiment

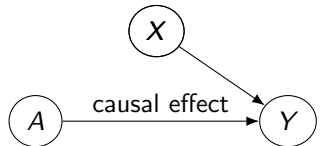
Observed data

- There are n units in the experiment
- For i-th unit, observed some covariates X_i prior to treatment assignment
- Binary treatment (exposure): $A_i = 1$ is the treatment and $A_i = 0$ the control
- After treatment assignment, we measure an outcome variable Y_i

Potential outcomes

- $Y_i(0)$, $Y_i(1)$, the potential outcomes for i-th unit under control and treatment
- $Y_i = Y_i(A_i)$ (SUTVA, stable unit treatment value assumption)
- *Fundamental problem of causal inference*: it is impossible to observe both $Y_i(0)$ and $Y_i(1)$

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	?	-3.7	1	-3.7
2	2.3	?	0	2.3
⋮	⋮	⋮	⋮	⋮



Randomization allows causal identification

Average treatment effect: ATE=$E[Y_i(1)] - E[Y_i(0)]$

- In RCT, we can compare $E[Y_i \mid A_i = 1]$ and $E[Y_i \mid A_i = 0]$
- In general, $E[Y_i \mid A_i = 1] = E[Y_i(1) \mid A_i = 1] \neq E[Y_i(1)]$
- By randomization in RCT, we have independence $A_i \perp (Y_i(0), Y_i(1))$, implying

$$E[Y_i \mid A_i = 1] = E[Y_i(1)].$$

Hence, in RCT,

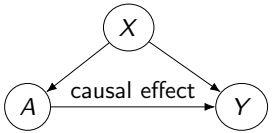
$$\text{ATE} = E[Y_i \mid A_i = 1] - E[Y_i \mid A_i = 0].$$

* What we just did is **causal identification** - writing counterfactuals using observed quantities.

* The rest is purely statistical and there are many available statistical inference methods, e.g., Ye et al. (2022).

Observational studies under no unmeasured confounding

- A confounder is a variable that influences both treatment assignment and the potential outcomes.
- Suppose all confounders are measured, which we denote as X. In other words, *people who look comparable are comparable*. This is called the no unmeasured confounders assumption (or strong ignorability, or conditional exchangeability).



- Then treatment among individuals with a particular value of X is essentially at random. So we find a "natural experiment" from observational data. Formally,

$$A \perp (Y(0), Y(1)) \mid X$$

- (Positivity or Overlap assumption) $0 < P(A = 1 \mid X) < 1$ for all X

Matching (without replacement)

- Matching is a method of sampling from a large reservoir of control units to produce a control group that is comparable to the treated group
- Matching is to design observational studies to resemble randomized experiments
- Advantages of matching:
  1. Transparent and intuitive, the balance can be displayed in a "Table 1" that is easily understood
  2. Clear display of lack of overlap in the covariate distributions between the treated and control group
  3. Model-based adjustment on matched samples is usually more robust to departures from the assumed form of the underlying model, primarily because of reduced reliance on the model's extrapolations
  4. Avoid researcher bias, as it enables blinding of the outcome information during matching

Propensity score subclassification and matching

- Practically, it is infeasible to match each treated subject to a control who appears nearly the same in terms of observed covariates. For instance, with 20 binary covariates, there are 2^{20} or about a million types of individuals
- Propensity score (Rosenbaum and Rubin, 1983) $\pi(X) = P(A = 1 \mid X)$ is a scalar and coarsest summary of the observed covariates X that can make the treated and control groups comparable, i.e.,

$$A \perp (Y(0), Y(1)) \mid \pi(X)$$

- In practice, we estimate $\pi(X)$, commonly by fitting a logistic regression of A_i on X_i. Denote the estimated propensity score as $\hat{\pi}(X_i)$.
- Two common approaches:
  1. Subclassification: we stratify by the propensity score, e.g. with five subclasses.
  2. Optimal matching: we form matched pairs of treated and control units so as to minimize the average within-pair difference in the propensity score

Multivariate matching (Rosenbaum, 2010, Chapters 8-13)

- A better matching procedure (superior in balancing the covariates, Rosenbaum and Rubin (1985)): multivariate matching with propensity score calipers
- Multivariate matching methods attempt to produce matched pairs or sets that balance observed covariates, so that, in aggregate, the distributions of observed covariates are similar in treated and matched control groups.

I. Distance for matching

- Within matched sets, we want subjects to be similar in terms of propensity scores and as close as possible in terms of covariates
- Our choice of distance to reflect these goals: robust Mahalanobis distance with a propensity score caliper.
- Mahalanobis distance: Let $\hat{\Sigma}$ be the sample covariance matrix of $\boldsymbol{x}$, then the estimated Mahalanobis distance between $\boldsymbol{x}_k$ and $\boldsymbol{x}_\ell$ is $(\boldsymbol{x}_k - \boldsymbol{x}_\ell)^T \hat{\Sigma}^{-1} (\boldsymbol{x}_k - \boldsymbol{x}_\ell)$.
  - Mahalanobis distance was originally developed for multivariate normal data. When the data are not normal, the Mahalanobis distance can exhibit some odd behavior.
- **Robust Mahalanobis distance**[1]
  1. replaces each of the covariates, one at a time, by its ranks, with average ranks for ties
  2. adjusts the covariance matrix of the ranks so it has a constant diagonal

[1] Step (i) limits the influence of outliers. Step (ii) prevents heavily tied covariates, such as rare binary variables, from having increased influence due to reduced variance.

Add propensity score caliper

- **Hard caliper**: With a caliper of width w, if two individuals, say k and ℓ, have logit propensity scores that differ by more than w, then the distance between these individuals is set to ∞. Matching with hard caliper may be infeasible.
- **Soft caliper**: we add a large but finite penalty (e.g., 1000) if the logit propensity scores are further apart than w
- Hansen (2011) suggests starting with a caliper of 50% of the pooled within-group standard deviation of the logit propensity score[2]. The caliper could be reduced if balance is not acceptable.

[2] The pooled within-group standard deviation is $\sqrt{s_1^2/2 + s_0^2/2}$, where s_1^2, s_0^2 are respectively the sample variance of logit propensity scores for the treated and control before matching.

A tiny example from genetic toxicology

- Welders are exposed to chromium and nickel, substances that can cause inappropriate links between DNA and proteins, which in turn may disrupt gene expression or interfere with replication of DNA.
- Costa et al. (1993) measured DNA-protein cross-links in samples of white blood cells from 21 railroad arc welders exposed to chromium and nickel and 26 unexposed controls.
- There are three covariates: age, race and current smoking behavior

Table 8.3 Estimated propensity scores $\widehat{e}(\mathbf{x})$ for 21 railroad arc welders and 26 potential controls. Covariates are age, race (C=Caucasian, AA=African American), current smoker (Y=yes, N=no).

Welders					Controls				
ID	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$	ID	Age	Race	Smoker	$\widehat{e}(\mathbf{x})$
1	38	C	N	0.46	1	48	AA	N	0.14
2	44	C	N	0.34	2	63	C	N	0.09
3	39	C	Y	0.57	3	44	C	Y	0.47
4	33	AA	Y	0.51	4	40	C	N	0.42
5	35	C	Y	0.65	5	50	C	N	0.23
6	39	C	Y	0.57	6	52	C	N	0.20
7	27	C	N	0.68	7	56	C	N	0.15
8	43	C	Y	0.49	8	47	C	N	0.28
9	39	C	Y	0.57	9	38	C	N	0.46
10	43	AA	N	0.20	10	34	C	N	0.54
11	41	C	Y	0.53	11	42	C	N	0.38
12	36	C	N	0.50	12	36	C	Y	0.64
13	35	C	N	0.52	13	41	C	N	0.40
14	37	C	N	0.48	14	41	AA	Y	0.35
15	39	C	Y	0.57	15	31	AA	Y	0.55
16	34	C	N	0.54	16	56	AA	Y	0.13
17	35	C	Y	0.65	17	51	AA	N	0.12
18	53	C	N	0.19	18	36	C	Y	0.64
19	38	C	Y	0.60	19	44	C	N	0.34
20	37	C	N	0.48	20	35	C	N	0.52
21	38	C	Y	0.60	21	34	C	Y	0.67
					22	39	C	Y	0.57
					23	45	C	N	0.32
					24	42	C	N	0.38
					25	30	C	N	0.63
					26	35	C	Y	0.65
	Mean Age	AA %	Smoker %	Mean $\widehat{e}(\mathbf{x})$		Mean Age	AA %	Smoker %	Mean $\widehat{e}(\mathbf{x})$
	38	10	52	0.51		43	19	35	0.39

Table 8.6 Rank-based Mahalanobis distances within propensity score calipers. Rows are the 21 welders and columns are for the first 6 of 26 potential controls. An ∞ signifies that the caliper is violated.

Welder	Control 1	Control 2	Control 3	Control 4	Control 5	Control 6
1	∞	∞	5.98	0.33	∞	∞
2	∞	∞	∞	0.47	∞	∞
3	∞	∞	∞	∞	∞	∞
4	∞	∞	10.43	∞	∞	∞
5	∞	∞	∞	∞	∞	∞
6	∞	∞	∞	∞	∞	∞
7	∞	∞	∞	∞	∞	∞
8	∞	∞	0.04	3.92	∞	∞
9	∞	∞	∞	∞	∞	∞
10	0.25	∞	∞	∞	3.72	4.01
11	∞	∞	0.28	∞	∞	∞
12	∞	∞	7.61	0.98	∞	∞
13	∞	∞	9.02	∞	∞	∞
14	∞	∞	6.83	0.64	∞	∞
15	∞	∞	∞	∞	∞	∞
16	∞	∞	10.61	∞	∞	∞
17	∞	∞	∞	∞	∞	∞
18	3.33	∞	∞	∞	0.05	0.01
19	∞	∞	∞	∞	∞	∞
20	∞	∞	6.83	0.64	∞	∞
21	∞	∞	∞	∞	∞	∞

II. Optimal matching

- Goal: match each treated subject with a control to minimize the total distance within matched pairs
- Greedy matching vs. optimal matching
  - Greedy matching: Choose treated-control pair with smallest distance, remove that treated and control pair, then choose remaining treated-control pair with smallest distance and so on until all treated units are paired.
  - Optimal matching: Efficiently achieve the above goal by solving an assignment problem (Rosenbaum, 1989)

	Control 1	Control 2
Treated 1	0	1
Treated 2	1	1000

- Hansen's R package optmatch implements optimal matching

III. Examining balance: three basic tools[3]

1. Standardized Differences

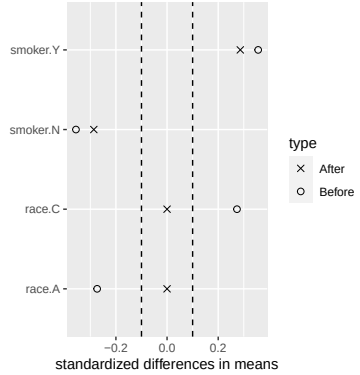
$$\text{stand. diff} = \frac{\bar{X}_{\text{trt}} - \bar{X}_{\text{matched ctl}}}{\sqrt{(s^2_{\text{trt}} + s^2_{\text{overall ctl}})/2}}$$

- We use $s^2_{\text{overall ctl}}$ to keep the denominator the same when comparing the stand. diff.
- Rule of thumb (Silber et al., 2001):
  Stand. diff. less than 0.1 is ideal;
  Stand. diff. between 0.1-0.2 are not ideal but acceptable;
  Stand. diff. greater than 0.2 indicate substantial imbalance.

[3] Other useful tools include the cross match test (Heller et al., 2010), classification permutation test (Gagnon-Bartsch and Shem-Tov, 2019)

III. Examining balance: three basic tools

2. Love Plot (invented by Thomas Love): Graphical way of showing how the standardized differences changed from before matching to after.



3. *t*-test for each variable

IV. Improving Match

- If a variable was out of balance, we could seek to improve its balance by adding a penalty for mismatches on the variable to the distance matrix. For example, we could add a penalty for mismatches on the variable.
- This may take some iterations and experience.
- Because outcomes are not available during this process, the search for a good match neither biases analyses of outcomes nor requires corrections for multiple inference.

V. Inference about treatment effect after match

In a pair match study in which we do not exclude any treated units, the treatment effect that we are estimating is the treatment on treated effect.

Inference methods

- Randomization tests (Lecture 1): e.g., Fisher exact test, Cochran–Mantel–Haenszel test, Wilcoxon's rank sum test, Wilcoxon's signed rank test, and permutation tests
- Fit a regression model controlling for the matched set indicators
  - Rubin (1973, 1979) found that the combined use of matching and model-based adjustments was both robust and efficient.
  - See also Guo and Rothenhäusler (2022).

Optimal matching with multiple controls

- In many cases, it is possible to match each treated to k control ($k > 1$)
- Substantial gains in efficiency are to be had using 2 controls and meaningful gains in efficiency are from 3-4 controls but for a much large number of controls, the gains are no longer large

Table 8.16 Variance ratio $1 + 1/k$ when matching k controls to each treated subject. Here, $k = \infty$ is only slightly better than $k = 6$, which is slightly better than $k = 4$.

Number of Controls	1	2	4	6	10	∞
Variance Multiplier	2	1.50	1.25	1.17	1.10	1

- Matching with multiple controls can potentially increase bias.
- Recommendation: We need to carefully check if we can achieve balance when we match with multiple controls, and if we cannot, we should resort back to matching with only one control.

Exact matching on a few key covariates

Exact matching

- For example, for studying treatment of cancer, we might want to match treated and control patients exactly on stage.
- When feasible, an exact match can be found by subdividing the matching problem into several smaller problems and piecing the answers together.
- This can also significantly reduce computation burden, especially for big datasets.

Almost exact matching

- Similar to the use of soft caliper, if subjects k and ℓ differ on these key covariates, then add a substantial penalty to the distance between them

Missing covariate values

A missing indicator procedure to balance the observed covariate and pattern of missingness:

1. Append missing indicator variable for each covariate with missing values
2. For each such variable, impute an arbitrary but fixed value
    * The presence of missing value indicators means that the arbitrary values matter little
3. Proceed with propensity score estimation and matching with these variables

Other more advanced matching techniques

- **Optimal full matching** (Rosenbaum, 2010, Chapter 8.6): Divides the sample into a collection of matched sets consisting either of a treated subject and any positive number of controls or a control unit and any positive number of treated units.
- **Fine balance**: (Rosenbaum, 2010, Chapter 10): Produce exact overall balance on a nominal covariate without matching exactly for that covariate. It can be used on top of the basic matching paradigm.
- **Optimal nonbipartite matching** (Rosenbaum, 2010, Chapter 11): Form pairs of individuals who look comparable but quite different in terms of doses of treatment

Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266.

Costa, M., Zhitkovich, A., and Toniolo, P. (1993). Dna-protein cross-links in welders: molecular implications. *Cancer research*, 53(3):460–463.

Dorn, H. F. (1953). Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health*, 43(6_Pt_1):677–683.

Gagnon-Bartsch, J. and Shem-Tov, Y. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics*, 13(3):1464–1483.

Guo, K. and Rothenhäusler, D. (2022). On the statistical role of inexact matching in observational studies. *Biometrika*, page asac066.

Hammond, E. C. (1964). Smoking in relation to mortality and morbidity. findings in first thirty-four months of follow-up in a prospective study started in 1959. *Journal of the National Cancer Institute*, 32(5):1161–1188.

Hansen, B. B. (2011). Propensity score matching to extract latent experiments from nonexperimental data: A case study. In *Looking Back*, pages 149–181. Springer.

Heller, R., Rosenbaum, P. R., and Small, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *The American Statistician*, 64(4):299–309.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Moertel, C. G., Fleming, T. R., Creagan, E. T., Rubin, J., O'Connell, M. J., and Ames, M. M. (1985). High-dose vitamin c versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: a randomized double-blind comparison. *New England Journal of Medicine*, 312(3):137–141.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

Rosenbaum, P. R. (2002). *Observational studies*. Springer.

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The annals of applied statistics*, 2(3):808–840.

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Even-Shoshan, O., Chen, W., Zhang, X., and Mosher, R. E. (2001). Multivariate matching and bias reduction in the surgical outcomes study. *Medical care*, pages 1048–1064.

Ye, T., Shao, J., Yi, Y., and Zhao, Q. (2022). Toward better practice of covariate adjustment in analyzing randomized clinical trials. *Journal of the American Statistical Association*, pages 1–13.