Chapter 18
Variable Selection for Causal Inference

Jihu Lee

August 17, 2021

Introduction

- *How to select the variables L for adjustment?*
- This chapter offers some guidelines for variable selection when the goal of the data analysis is Causal Inference
- Goals, Variables' effect on bias, Machine learning, Doubly robust estimators

Table of Contents

Different goals of variable selection

- Predictive/Associational models
  - may want to select any variables that improve predictive ability
  - adjustment for confounding is unnecessary
  - automated algorithms: *lasso*,$\cdots$
- Causal models
  - thoughtful selection of confounders is needed
  - usually require adjustment for confounding and other biases

Table of Contents

Variables that induce or amplify bias

- Ideal situation
  - unlimited computational power, dataset with a quasi-infinite number of individuals, many variables
- Some of variables may be confounders - want to adjust
- Unbiased estimate the average causal effect $E[Y^{a=1}] - E[Y^{a=0}]$
- Want to ensure that some variables are not selected for adjustment because adjustment for those variables would induce bias

Variables that induce bias

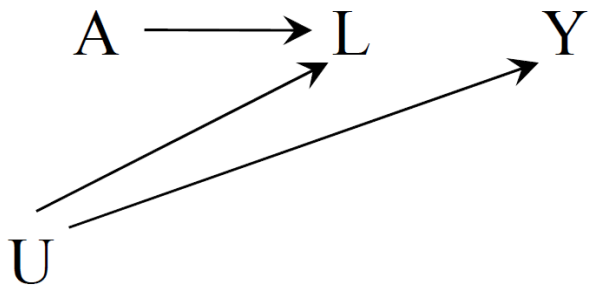

Figure 18.1

- $E[Y^{a=1}] - E[Y^{a=0}] = 0$ is unbiasedly estimated by $E[Y|A = 1] - E[Y|A = 0]$ (no confounding)
- adjustment for L by g-formula

$$\sum_l E[Y|A = 1, L = l]Pr(L = l) - \sum_l E[Y|A = 0, L = l]Pr(L = l)$$

- $Pr(L = l) \neq Pr(L = l|A) \rightarrow$ biased

Variables that induce bias

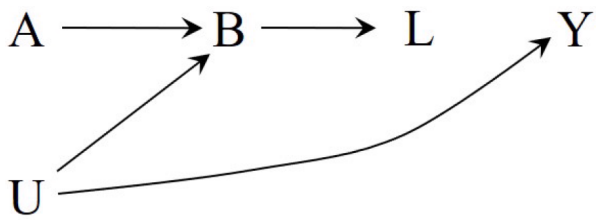

Figure 18.2

- same bias is expected to arise when we adjust for L
- *selection bias under the null*

Variables that induce bias

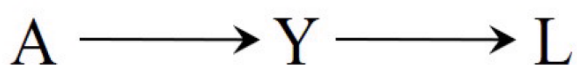

Figure 18.3

- $E[Y^{a=1}] - E[Y^{a=0}] \neq 0$ is unbiasedly estimated by $E[Y|A = 1] - E[Y|A = 0]$ (no confounding)
- induce bias for same reasons
- if $A \rightarrow Y$ absent: g-formula contrast would be zero & unbiased
- *selection bias under the alternative*

Variables that induce bias

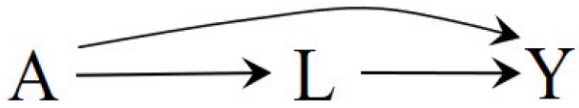

Figure 18.4

- *overadjustment for mediators*

Variables that induce bias

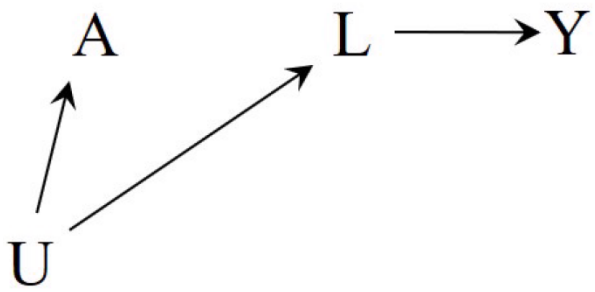

Figure 18.5

- not adjusting post-treatment variables can solve the problem?
- causal graphs do not care about temporal order

Variables that induce bias

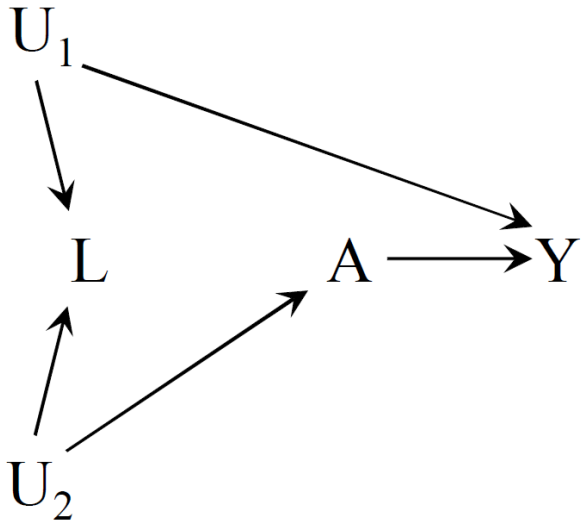

Figure 18.6

- L: pre-treatment, collider/confounder (cannot distinguish)
- adjusting it will introduce M-bias(selection bias)
- must rely on external information to decide whether to adjust

Variables that amplify bias

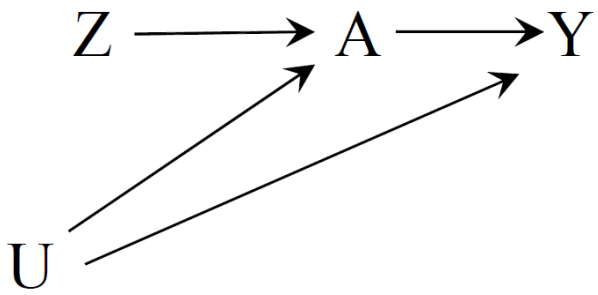

Figure 18.7

- U: not available in the data, cannot adjust, confounding is intractable
- Z: adjustment does not eliminate confounding due to U, instrument
- could amplify/reduce bias (unknown)

Table of Contents

Causal inference and machine learning

- assumption) no variables that may induce or amplify bias
- Standardization: estimate the mean Y cond. on X, $b(X)$
- IP-weighting: estimate the prob. of A cond. on X, $\pi(X)$
- produce $\hat{b}(x)$, $\hat{\pi}(x)$ via parametric models

Causal inference and machine learning

- when handling high-dimensional problems: *lasso*,$\cdots$
- Machine learning algorithms do not guarantee that the selected variables will eliminate confounding: use doubly robust est.
- Machine learning algorithms are statistical black boxes: unbiasedness$\neq$correct variance

Table of Contents

Doubly robust machine learning estimators

- To construct valid confidence intervals with small bias
  1. Sample splitting
  2. Cross-fitting

Sample splitting

1. randomly divide the study population into two halves: estimation sample / training sample
2. apply predictive algorithms to the training sample: obtain $\hat{b}(x)$, $\hat{\pi}(x)$
3. compute doubly robust estimator in the estimation sample

- allows to use standard statistical inference procedures based on half of individuals

Cross-fitting

1. repeat sample splitting procedure 2 3 but swapping the roles of the halves
2. compute doubly robust estimator in the new estimation sample
3. compute the average of two estimators

- statistical properties & use all the data
- detect whether the bias is too large: active reasearch

Table of Contents

Variable selection is a difficult problem

- available subject-matter knowledge may be insufficient to identify all important confounders
- no machine learning algorithm is optimal in all settings
- implementation of doubly robust estimators is difficult-computationally expensive
- no guarantee that the variance of the causal effect gained from doubly robust estimators will be small enough