程序代写代做 CS编程辅导

Project 8
Super... ...argeted Maximum Likelihood
...stimation
...Computational Social Science

WeChat: cstutorcs

1 Introduction

Heart disease is the leading cause of death in the United States and treating it properly is an important public health goal. However, it is a complex disease with several different risk factors and potential treatments. Physicians typically recommend changes in diet, increased exercise, and/or medication to treat symptoms, but it is difficult to determine how effective any one of these factors is in treating the disease. In this project, you will explore SuperLearner, Targeted Maximum Likelihood Estimation (TMLE), and Longitudinal Targeted Maximum Likelihood Estimation (LTMLE). Using a simulated dataset, you will explore whether taking blood pressure medication reduces mortality risk.

2 Data

This dataset was simulated using R (so it does not come from a previous study or other data source). It contains several variables:

- **blood_pressure_medication**: Treatment indicator for whether the individual took blood pressure medication (0 for control, 1 for treatment)

- **mortality**: Outcome indicator for whether the individual passed away from complications of heart disease (0 for no, 1 for yes)

- **age**: Age at time 1

- **sex_at_birth**: Sex assigned at birth (0 female, 1 male)

- **simplified_race**: Simplified racial category. (1: White/Caucasian, 2: Black/African American, 3: Latinx, 4: Asian American, 5: Mixed Race/Other)

- **income_thousands**: Household income in thousands of dollars

- **college_educ**: Indicator for college education (0 for no, 1 for yes)
- **bmi**: (BMI)
- **cho**:
- **blo**: blood pressure
- **bmi**: time 2
- **cho**: red at time 2
- **blood_pressure_2**: BP measured at time 2
- **blood_pressure_medication_2**: Whether the person took treatment at time period 2

For the "SuperLearner" and "TMLE" portions, you can ignore any variable that ends in "_2", we will reintroduce these for LTMLE.

3 SuperLearner

3.1 Modeling

Fit a SuperLearner model to estimate the probability of someone dying from complications of heart disease, conditional on treatment and the relevant covariates. Do the following:

1. Choose a library of at least 3 machine learning algorithms to evaluate. **Note**: We did not cover how to hyperparameter tune constituent algorithms within SuperLearner in lab, but you are free to do so if you like (though not required to for this exercise).

2. Split your data into train and test sets.

3. Train SuperLearner

4. Report the risk and coefficient associated with each model, and the performance of the discrete winner and SuperLearner ensemble

5. Create a confusion matrix and report your overall accuracy, recall, and precision

3.2 Discussion Questions

1. Why should we, in general, prefer the SuperLearner ensemble to the discrete winner in cross-validation? Or in other words, what is the advantage of "blending" algorithms together and giving them each weights, rather than just using the single best algorithm (with best being defined as minimizing risk)?

4 Targeted Maximum Likelihood Estimation

4.1 C

TMLE re   models:

1. The   e relationship between the outcome and the
   trea (A, W).

2. The   el, or the relationship between assignment to
   treatment and predictors $P(A|W)$

Using ggdag and daggity, draw a directed acylcic graph (DAG) that describes
the relationships between the outcome, treatment, and covariates/predictors.
Note, if you think there are covariates that are not related to other variables
in the dataset, note this by either including them as freestanding nodes or by
omitting them and noting omissions in your discussion.

4.2 TMLE Estimation

Use the 'tmle' package to estimate a model for the effect of blood pressure
medication on the probability of mortality. Do the following:

1. Use the same SuperLearner library you defined earlier

2. Use the same outcome model and propensity score model that you spec-
   ified in the DAG above. If in your DAG you concluded that it is not
   possible to make a causal inference from this dataset, specify a simpler
   model and note your assumptions for this step.

3. Report the average treatment effect and any other relevant statistics

4.3 Discussion Questions

1. What is a "double robust" estimator? Why does it provide a guarantee
   of consistency if either the outcome model or propensity score model is
   correctly specified? Or in other words, why does mispecifying one of the
   models not break the analysis? **Hint**: When answering this question,
   think about how your introductory statistics courses emphasized using
   theory to determine the correct outcome model, and in this course how
   we explored the benefits of matching.

5 LTMLE Estimation

Now imagine that everything you measured up until now was in "time period
1". Some people either choose not to or otherwise lack access to medication in
that time period, but do start taking the medication in time period 2. Imagine
we measure covariates like BMI, blood pressure, and cholesterol at that time for
everyone in the study.

5.1 Causal Diagram

Update y... ...ncorporate this new information.

Hint: ...m Maya's lecture, or slides 15-17 from Dave's second sli... ...ecture on Propensity Score Weighting in Week 9.

Hint: ... ...y of the variables that end in "_2" are likely affected b... ...variates and the first treatment when drawing your DAG...

5.2 LTMLE Estimation

Use the 'ltmle' package for this section. First fit a "naive model" that **does not** control for the time-dependent confounding. Then run a LTMLE model that does control for any time dependent confounding. Follow the same steps as in the TMLE section. Do you see a difference between the two estimates?

5.3 Discussion Questions

1. What sorts of time-dependent confounding should we be especially worried about? For instance, would we be concerned about a running variable for age the same way we might be concerned about blood pressure measured at two different times?