Observational Studies Simulation Homework

Andrea Cornejo, Ray Lu & Zarni Htet

Objective

The goal of this exercise is to learn how to simulate a few different types of observational causal structures and evaluate the properties of different approaches to estimating the treatment effect through linear regression.

Problem Statement

You should be familiar with the assumptions of linear regression (both **structural** and **parametric**) for causal effect estimation. Suppose we want to simulate a simple causal data set from the joint distribution of the covariates, treatment, and potential outcomes.

The data generating process (DGP) is:p(X, Z, Y0,Y1)=p(X)p(Z|X)p(Y1,Y0|Z, X). (As per usual, X is the pretest variable, Z is the treatment variable and Y0 and Y1 are the potential outcomes.)

Part A: Linear Parametric form

Question 1: Simulate the data

(a) Start with the marginal distribution of X. Simulate as X~N(0,1) with sample size of 1000. Set the seed to be 1234.

(b) Look at the DGP. What role does X play?

(c) The distribution of binary Z depends on the value of X. Therefore, the next step is to simulate Z from p(Z|X) = Binomial(p), where the vector of probabilities can vary across observations. Come up with a strategy for generating the vector Z conditional on X that forces you to create be explicit about how these probabilities are conditional on X (an inverse logit function would be one strategy but there are others). Make sure that X is significantly associated with Z and that the vector of probabilities used to draw Z doesn't vary below .05 or above .95.

(d) The last step is to simulate Y from p(Y0,Y1|Z,X). Come up with a strategy for simulating each potential outcome with appropriate conditioning on Z and X with the following stipulations.

- (i) Make sure that E[Y(1)|X] - E[Y(0)|X] = 5.
- (ii) Make sure that X has a linear and statistically significant relationship with the outcome.
- (iii) Finally, set your error term to have a standard deviation of 1 and allow the residual standard error to be different for the same person across potential outcomes.
- (iv) Create a data frame containing X,Y,Y0,Y1 and Z and save it for use later.

(e) Think about the difference between the DGP used in this homework and the first DGP from previous homework (completely randomized experiment). How is the difference in the study design encoded?

(f) Calculate the SATE from 1.d.iv (save it for use later).

Question 2: Playing the role of the researcher

Now switch to the role of the researcher for a moment. Pretend someone handed you a dataset generated as specified above and asked you to estimate a treatment effect – for this you will use the dataset generated in 1f above. You will try two approaches: difference in means and regression.

(a) Estimate the treatment effect using a difference in mean outcomes across treatment groups (save it for use later).

(b) Estimate the treatment effect using a regression of the outcome on the treatment indicator and covariate (save it for use later).

(c) Create a scatter plot of X versus the observed outcome with different colors for treatment and control observations (suggested: red for treated and blue for control). If you were the researcher would be comfortable using linear regression in this setting?

Question 3: Exploring the properties of estimators

Now we're back to the role of god of Statistics.

(a) Create a scatter plot of X versus each potential outcome with different colors for treatment and control observations (suggested: red for Y(1) and blue for Y(0)). Is linear regression a reasonable model to estimate causal effects for the observed data set? Why or why not?

(b) Find the bias of each of the **estimates** calculated by the researcher in Question 2 relative to SATE.

(c) Think harder about the practical significance of the bias by dividing this estimate by the standard deviation of the observed outcome Y.

(d) Find the bias of each of the **estimators** by creating randomization distributions for each. [Hint: When creating randomization distributions remember to be careful to keep the original sample the same and only varying treatment assignent and the observed outcome.]

Part B: Non-Linear Parametric form

Now we'll explore what happens if we fit the wrong model in an observational study.

Question 1: Simulate the data

(a) Create function sim.nlin with the following DGP.

- (i) X should be drawn from a uniform distribution between 0 and 2.
- (ii) Treatment assignment should be drawn from a Binomial distribution with the following properities (make sure you save the p vector for use later).

$$\mathrm{E}[Z \mid X] = p = \mathrm{logit}^{-1}(-2 + X^2) Z \sim \mathrm{Binom}(N, p)$$

- (iii) The response surface (model for Y(0) and Y(1)) should be drawn from the following distributions:

$$Y(0) = \mathbf{2}X + \epsilon_{\mathbf{0}}$$
$$Y(1) = \mathbf{2}X + \mathbf{3}X^2 + \epsilon_{\mathbf{1}}$$

where both error terms are normally distributed with mean 0 and standard deviation of 1.

- (iv) Make sure the returned dataset has a column for the probability of treatment assignment as well.

(b) Simulate a data set called data.nlin with sample size 1000.

(c) Make the following plots.

- (i) Create overlaid histograms of the probability of assignment.
- (ii) Make a scatter plot of X versus the observed outcomes versus X with different colors for each treatment group.
- (iii) Create a scatter plot of X versus each potential outcome with different colors for treatment and control observations (suggested: red for Y(1) and blue for Y(0)). Does linear regression of Y ond X seem like a good model for this response surface?

(d) Create randomization distributions to investigate the properties of each of 3 estimators with respect to SATE: (1) difference in means, (2) linear regression of the outcome on the treatment indicator and X, (3) linear regression of the outcome on the treatment indicator, X, and X^2.

(e) Calculate the standardized bias (bias divided by the standard deviation of Y) of these estimators relative to SATE.

PART C: OPTIONAL CHALLENGE QUESTION

Simulate Linear Causal Structure With Mutiple Covariates

(a) Simulate observational data set from following distribution P(X1,X2,X3,Y1,Y0,Z)=P(X1,X2,X3)×P(Z|X1,X2,X3) ×P(Y1,Y0|Z,X1,X2,X3).

Once again make sure that the probability of being treated for each person falls between .05 and .95 and there is a reasonable amount of overlap across the treatment and control groups. Generate the response surface as in the following:

$$Y(0) = X1 + X2 + X3 + \epsilon$$

$$Y(1) = X1 + X2 + X3 + 5 + \epsilon$$

(b) If you didn't want the covariates to be independent of each other, how could you simulate X1,X2 and X3 ?

(c) Create randomization distributions for (1) a regression estimator that controls for only one of the 3 covariates and (2) a regression estimator that controls for all 3 covariates. Evaluate the standardized bias of these estimators relative to SATE.