Homework Assignment 3

Econ 31380 Causal Machine Learning
Max H. Farrell

Due November 29. Submit your answers on Canvas typed using LaTeX or Markdown.

1 Propensity Score Weighting & ATT Estimation

This is a continuation from homework 2.

Assume that the random variables $(Y_1, Y_0, T, \mathbf{X}')' \in \mathbb{R} \times \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$ obey $\{Y_1, Y_0\} \perp\!\!\!\perp T \mid \mathbf{X}$. The researcher observes $(Y, T, \mathbf{X}')'$, where $Y = Y_1 T + Y_0(1-T)$. Define the propensity score $p(\mathbf{x}) = \mathbb{P}[T = 1 \mid \mathbf{X} = \mathbf{x}]$ and assume it is bounded inside $(0,1)$. Define $\mu_t = \mathbb{E}[Y(t) \mid T = 1]$ and $\mu_t(\mathbf{x}) = \mathbb{E}[Y(t) \mid \mathbf{X} = \mathbf{x}]$. The average treatment effect on the treated (ATT) is $\tau = \mu_1 - \mu_0$.

Assume that the propensity score is correctly specified as a logistic regression: for a d vector θ_0, it holds that $p(\mathbf{x}) = (1 + \exp\{-\theta_0'\mathbf{x}\})^{-1}$.

(a) Consider estimating θ_0 using maximum likelihood, denote the estimator $\widehat{\theta}_{\text{MLE}}$. Write down the objective function that is solved by the estimator and the equations that characterize the solution.

(b) Derive the influence function for $\widehat{\theta}_{\text{MLE}}$.

(c) Consider estimating θ_0 using nonlinear least squares, denote the estimator $\widehat{\theta}_{\text{NLS}}$. Write down the objective function that is solved by the estimator and the equations that characterize the solution.

(d) Derive the influence function for $\widehat{\theta}_{\text{MLE}}$. Compare it to the one for $\widehat{\theta}_{\text{NLS}}$.

Now we turn to ATT estimation and inference. Combining the moment conditions (see homework 2), the ATT obeys

$$\tau = \mu_1 - \mu_0 = \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 1] = \mathbb{E}\left[\frac{TY}{\mathbb{E}[T]}\right] - \frac{1}{\mathbb{E}[T]}\mathbb{E}\left[\frac{(1-T)p(\mathbf{X})Y}{(1-p(\mathbf{X}))}\right].$$

For an estimator $\hat{p}(\mathbf{x})$ of the propensity score, we will estimate the ATT using the sample analogue of the above moment condition. Let $\hat{p} = \sum_{i=1}^n t_i/n$ and define the estimator

$$\widehat{\tau} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^n \frac{t_i y_i}{\hat{p}} - \frac{1}{\hat{p}}\frac{1}{n}\sum_{i=1}^n \frac{(1-t_i)\hat{p}(\mathbf{x}_i)y_i}{(1-\hat{p}(\mathbf{x}_i))}$$

(e) Derive the influence function of your estimator assuming that you use maximum likelihood to estimate the propensity score.

(f) Derive the influence function of your estimator assuming that you use nonlinear least squares to estimate the propensity score.

(g) Conduct a simulation study where you use both first step estimation methods. Your study should verify the derivations above as well as compare the two estimators. Which performs better? Explore different sample sizes, dimensions of $\mathbf{X}$, noise levels, etc, i.e. vary different aspects of the simulation design.

2 Nonparametric Density Estimation

(Density estimation isn't as useful as nonparametric regression, in general and for casual inference in particular, but all the conceptual lessons learned here carry over to regression.)

We have an i.i.d. sample $\{x_1, \ldots, x_n\}$ from a scalar random variable $X \in R$, where X has the cdf $F(x)$ and the (Lebesgue) density $f(x)$. Assume X has compact, connected support and that $f(x)$ is bounded and bounded away from zero. Our goal in this problem is to learn $F(x)$ and $f(x)$ at a single point x.[1]

(a) Consider the empirical distribution function

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\}. \tag{1}$$

Motive this estimator as the sample analogue of the population cdf. Prove that $\hat{F}(x)$ is unbiased and compute its variance. Establish that the estimator is consistent.

(b) Prove, including providing sufficient conditions, that $\sqrt{n}(\hat{F}(x) - F(x)) \to_d \mathcal{N}(0, \Omega)$. Characterize the variance Ω and provide a consistent estimator.

(c) Suppose that you know that $X \sim \mathcal{N}(\mu, \sigma^2)$. Use the sample mean and variance to provide an estimator of the cdf, call it $\tilde{F}(x)$. Prove that this estimator is consistent and asymptotically Normal.

(d) Conduct a simulation study to examine the empirical performance of both $\hat{F}(x)$ and $\tilde{F}(x)$. Evaluate the consistency and the variance (i.e. the CLT) for both estimators. If the true distribution is Normal, which is more efficient? What happens when the distribution is not Normal? Try several different distributions as well as different parameters for those distributions. Choose three representative values x at which to study $F(x)$. Study what happens as n changes.

Now we turn to estimating the density $f(x)$. The density is the derivative of the cdf, and therefore is given by

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}. \tag{2}$$

(e) Use (1) and (2), for a fixed h, to give a plug in estimator for $f(x)$ denoted $\hat{f}(x)$.

(f) For fixed h, compute the bias of $\hat{f}(x)$. Prove that the bias vanishes as $h \to 0$.

(g) Assume that $f(x)$ is twice continuously differentiable. Prove that the bias of $\hat{f}(x)$ is $O(h)$ and characterize the constant. That is, show that

$$\mathbb{E}[\hat{f}(x) - f(x)] = Kh + o(h)$$

and give the precise form of K.

(h) For fixed h, compute the variance denoted $\Sigma = \mathbb{V}[\hat{f}(x)]$. Provide a consistent estimator.

(i) Compute the mean square error of your estimator and find the value of h that minimizes it. Characterize precisely what happens to this optimal h as $n \to \infty$. How would you choose h in an application for the goal of estimation?

[1] What we would ideally have is uniform estimation and inference, that holds for all x simultaneously. This is harder to achieve, but can be done. See class discussion of `binsreg`.

(j) For fixed h, prove that

$$\frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\Sigma^{1/2}} \to_d \mathcal{N}(0,1).$$

(k) Provide sufficient conditions so that

$$\frac{\hat{f}(x) - f(x)}{\Sigma^{1/2}} \to_d \mathcal{N}(0,1).$$

Characterize precisely the requirements that h must obey as $n \to \infty$.

(l) Compare the requirements on h in part **(k)** to what you found in part **(i)**. Discuss what you find. How would you choose h in an application for the goal of inference?

(m) Conduct a simulation study to examine the empirical performance of $\hat{f}(x)$. Evaluate the bias and variance of your estimator and the quality of the Normal approximation. Compute the empirical coverage and length of 95% confidence intervals. Study what happens as you vary n, h, the true distribution, and the evluation point x.

3 Appplication

The file `Banerji-Berry-Shotland_2017_AEJ.csv` contains data from a recent paper.[2] The outcome is a (normalized) child's test, in `caser_total_norm`. `treatment` has four different values, indicating different trainings for mothers. The first is the baseline/control. There are six X variables (dummies) and three W variables (continuous).[3] We want to explore the impact of each treatment relative to the baseline (`treatment`=1).

3.1 LASSO & Discrete Heterogeneity

(a) Run a single linear regression that provides estimates and inference for $\mu_t = \mathbb{E}[Y(t)]$, $t = 1, 2, 3, 4$. Add covariates to the regression to see if efficiency is improved. First add the covariates directly and then do it demeaned and interacted. Try adding interactions among the X and W.

(b) Use the LASSO to select controls in one of the models you ran above. Leave the treatment coefficients unpenalized. Is precision improved?

(c) Choose one of the X variables out of the six. Run a single linear regression that provides estimates and inference for the heterogeneous effects $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$ for $x = \{0, 1\}$ (i.e. eight total numbers).

(d) Add all the other X variables, and the W variables, and interactions and polynomials, and apply the lasso to select controls while still giving inference on the eight $\mu_t(x)$. Is precision improved?

[2] This isn't the full data. I took a clean subset for illustration. The paper is "The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India" by Banerji, Berry, Shotland (2017, *American Economic Journal: Applied Economics*), available here: `https://www.aeaweb.org/articles?id=10.1257/app.20150390`.

[3] The X variables are (i) if the child is male, (ii) if the state is Bihar, (iii) if the mother has any education, (iv) if the father has any education, (v) if the mother is over 30 years old, (vi) if the family income is from farming. The W variables are These are child's age, number of kids in the household, and a baseline test score.

(e) Split the data randomly in two pieces, call them sample A and sample B. In sample A, use the lasso to identify the most impacted subgroups based on X and interactions in X (go up to only two- or three-way interactions). Use sample B to validate the size of these impacts and do hypothesis testing. Discuss the role played by sample splitting in this case.

3.2 Binsreg & Continuous Heterogeneity

For $j = 1, 2, 3$, define $\omega_t(w_j) = \mathbb{E}[Y(t) \mid W_j = w_j]$.

(f) Use `binsreg` to plot all possible $\omega_t(w_j)$ (probably not in one picture). What did you specify for the other controls and why?

(g) Pick one W_j and use confidence bands to assess a substantive question about $\omega_t(w_j)$, $t = 1, 2, 3, 4$. For example, is it monotonic? Are there decreasing returns? Etc.

3.3 Deep Nets and Forests

(h) Consider the model

$$Y_i = \sum_{t=1}^{4} \mu_t(x, w) \mathbb{1}\{T_i = t\} + \varepsilon_i.$$

Provide conditions under which the functions $\mu_t(x, w) = \mathbb{E}[Y(t) \mid X = x, W = w]$ are identified.

(i) Apply random forests to learn $\mu_t(x, w)$ full flexibly. For each one, create a partial dependence plot for each continuous w_j. How do these compare to what you found in **(f)**? For each $\mu_t(x, w)$, create and discuss the variable importance plot. Do these make sense to you for this application?

(j) Use neural networks to learn $\mu_t(x, w)$ full flexibly. Try several different architectures for your deep nets. Select a single one as the best and justify your choice.

(k) Conduct inference on the treatment effect of treatment t compared to baseline, $\mathbb{E}[\mu_t(X, W) - \mu_0(X, W)]$, using the influence function based estimation from class and preliminary estimates from both **(i)** and **(j)**.