ECMA 31380 - Causal Machine Learning - Homework 3

Fernando Rocha Urbano

Autumn 2024

Attention: all code is available in

`https://github.com/Fernando-Urbano/causal-machine-learning/tree/main/hw4`.

1 Conditions on Nonparametric Estimators

We are studying the impact of a multi-valued treatment $T \in \{0, 1, \ldots, T\}$, for some integer T, on an outcome Y. We observe $Z = (Y, T, \mathbf{X}')' \in \mathbb{R} \times \{0, 1, \ldots, T\} \times \mathbb{R}^d$. Define the potential outcomes as $Y(t)$, the propensity score $p_t(\mathbf{x}) := \mathbb{P}[T = t \mid \mathbf{X} = \mathbf{x}]$, and the regression functions $\mu_t(x) = \mathbb{E}[Y(t) \mid \mathbf{X} = \mathbf{x}]$.

Interesting estimands can be built from averages of $\mu_t(\mathbf{x})$. For example: the ATE of treatment level t is $\tau_t = \mathbb{E}[\mu_t(\mathbf{X}) - \mu_0(\mathbf{X})]$. If the treatment is a dose, then the effect of increasing the dose from t to $t + 1$ is $\mathbb{E}[\mu_{t+1}(\mathbf{X}) - \mu_t(\mathbf{X})]$. And so on. So we will study $\mu_t = \mathbb{E}[\mu_t(\mathbf{X})] = \mathbb{E}[Y(t)]$.

Suppose that $p_t(x) = p_t$ is constant over x, so that this is a randomized experiment.

1.a Linear Regression Model and Assumptions

Provide a single linear regression model that yields identification of all μ_t, $t \in \{0, \ldots, T\}$. What assumptions do you need? Describe the estimators $\hat{\mu}$. Provide regularity conditions so that the vector $\hat{\mu}$ is asymptotically Normal, asymptotically unbiased, and characterize the asymptotic variance.

1.b Sufficient Conditions for Identification

Now suppose that $p_t(x)$ is not constant. Provide sufficient conditions so that μ_t is identified. Compare these conditions to what you found above.

In class, we studied nonparametric regression using piecewise polynomials of degree p (fixed) and $J = J_n \to \infty$ pieces and proved that the L_2 convergence rate is (using the notation of the present context)

$$\|\hat{\mu}_t(\mathbf{x}) - \mu_t(\mathbf{x})\|_2^2 = O_p\left(\frac{J^d}{n} + J^{-2(p+1)}\right).$$

Let the number of bins be $J = Cn^\gamma$ for some constants (positive) C and γ. We will ignore C and focus on rates here.
First we study nonparametric estimation and inference.

1.c Range of γ for Consistency

For what range of γ is $\hat{\mu}_t(\mathbf{x})$ consistent? How does this range depend on the dimension and the polynomial order? Are there values of p and d such that this interval is empty?

1.d Optimal Value of γ

What value of γ is optimal in the sense that the rate is the fastest? Call this $\gamma^\star_{\text{mse}}$. How does $\gamma^\star_{\text{mse}}$ vary with the dimension and the polynomial order?

1.e Asymptotic Normality of $\hat{\mu}_t(x)$

For what range of γ is $\hat{\mu}_t(x)$ asymptotically Normal when properly centered and scaled? That is, determine the range for γ such that

$$\sqrt{n/J^d}(\hat{\mu}_t(\mathbf{x}) - \mu_t(\mathbf{x})) \xrightarrow{d} \mathcal{N}(0, V).$$

(Don't worry about quantifying V). How does this range depend on the dimension and the polynomial order? Are there values of p and d such that this interval is empty?

1.f Range of γ for Optimal Rate $\gamma^\star_{\text{mse}}$

Is $\gamma^\star_{\text{mse}}$ in this range?

1.g Semiparametric Estimation and Inference

In class, we showed that there was a problem with the two-step plug-in estimator $\tilde{\mu}_t = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(x_i)$ and that it did not have the same influence function as the parametric regression-based plug-in estimator. However, Cattaneo and Farrell (2011) showed that it does in fact obtain an influence function representation, with the familiar influence function. That paper shows that if

$$\sqrt{n}\left(\frac{J^d}{n} + J^{-(p+1)}\right) \to 0$$

then

$$\sqrt{n}(\tilde{\mu}_t - \mathbb{E}[Y(t)]) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_i + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\psi_t(Z)^2]),$$

where $\psi_t(z_i) = \mu(x_i) - \mathbb{E}[Y(t)] + \mathbb{I}\{t_i = t\}(y_i - \mu(x_i))/p_t(x_i)$.

(i) For what range of γ is inference on the $\mathbb{E}[Y(t)]$ possible? How does this range depend on p and d? Are there values of p and d such that this interval is empty?

(ii) Is $\gamma^{\star}_{\text{mse}}$ in this range?

1.h Influence Function-Based Estimator

Now consider the influence function-based estimator. Let $\hat{\mu}_t(x)$ and $\hat{p}_t(x)$ be partitioning-based estimators of the respective functions, which both have the rate of Equation (1). Define

$$\hat{\mu}_t = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_t(x_i) + \frac{\mathbb{I}\{t_i = t\}(y_i - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)}\right\}.$$

In class, we proved that the linear representation and asymptotic normality of Equation (3) holds (with $\tilde{\mu}_t$ replaced by $\hat{\mu}_t$) if

$$\|\hat{\mu}_t(x) - \mu_t(x)\|_2 \to 0, \quad \|\hat{p}_t(x) - p_t(x)\|_2 \to 0, \quad \text{and} \quad \sqrt{n}\|\hat{\mu}_t(x) - \mu_t(x)\|_2\|\hat{p}_t(x) - p_t(x)\|_2 \to 0.$$

(i) For what range of γ is inference on the $\mathbb{E}[Y(t)]$ possible? How does this range depend on p and d? Are there values of p and d such that this interval is empty?

(ii) Is $\gamma^{\star}_{\text{mse}}$ in this range?

(iii) In terms of the allowed γ, compare your findings to the previous part.

2 Propensity Score Weighting & ATT Estimation

This is a continuation from homeworks 2 & 3.

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

In homework 3, you studied a "plug-in" estimator of the ATT given by

$$\hat{\tau}_{\text{PI}} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{t_i y_i}{\hat{p}} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-t_i)\hat{p}(x_i) y_i}{(1-\hat{p}(x_i))}.$$

In homework 2, you proved that

$$\mu_0 = \frac{1}{\mathbb{E}[T]}\mathbb{E}\left[T\mu_0(X) + \frac{(1-T)p(X)(Y-\mu_0(X))}{(1-p(X))}\right]$$

and that this moment condition is doubly robust. This motivates a doubly robust estimator of the ATT given by

$$\hat{\tau}_{\text{DR}} = \hat{\mu}_1 - \hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{t_i y_i}{\hat{p}} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-t_i)\hat{p}(x_i) y_i}{(1-\hat{p}(x_i))}\right\}.$$

We will conduct a simulation study to examine various properties of these estimators. Make sure your simulation study obeys the data-generating process assumptions, including overlap. In this case, we know from theory that cross-fitting is not necessary, so we'll skip it unless specifically asked for.

2.a High-Dimensional Parametric Case

(a) First, we study the high-dimensional parametric case. Suppose that $\mu_0(x) = \beta_0' x$ and $p(x) = (1 + \exp\{-\theta_0' x\})^{-1}$. Use a penalized linear model for $\hat{\mu}_0(x_i)$ and a penalized logistic regression for $\hat{p}(x)$. Try both LASSO and ridge regression.

Find the sampling distribution of both estimators (4) and (5) as the data-generating process varies. In particular, try all combinations of the following:

- Sample size $n = 1000$ and 5000,
- Dimension $d = \dim(x) = \{10, 50, 500, 5000\}$, and
- Sparsity levels $s_\beta = \|\beta_0\|_0 = \{d/10, d/2, d\}$ and $s_0 = \|\theta_0\|_0 = \{d/10, d/2, d\}$.

(i) What happens as n grows but d, s_β, s_0 are fixed?

(ii) What happens to $\hat{\tau}_{\mathrm{PI}}$ as d and s_0 change for fixed n?

(iii) How does s_β impact $\hat{\tau}_{\mathrm{PI}}$?

(iv) Verify the doubly robust property of $\hat{\tau}_{\mathrm{DR}}$.

(v) What happens if you do not penalize in the first stage, but just use plain OLS and logistic regression?

(vi) Discuss what your results mean for applied practice. When would you recommend the different estimators and why?

2.b Nonparametrics and Low-Dimensional Case

(b) Now we turn to nonparametrics and lower-dimensional functions. Suppose that $\mu_0(x)$ and $p(x)$ are completely unknown functions. In your data-generating process, make them nonlinear functions of x^2. Try $n = \{1000, 5000, 15000\}$ and $d = \dim(x) = \{1, 3, 5, 10\}$, including designs with sparsity. Use deep nets and random forests (and anything else you care to try).

(i) What happens as n grows but d is fixed?

(ii) Verify the doubly robus property of $\hat{\tau}_{\text{DR}}$.

(iii) Dicuss what your results mena for applied practice. When would you recommend the different estimators and why?

(iv) Verify the doubly robust property of $\hat{\tau}_{\text{DR}}$.

(v) What happens if you do not penalize in the first stage, but just use plain OLS and logistic regression?

(vi) Discuss what your results mean for applied practice. When would you recommend the different estimators and why?

Now real data. Return to the Census data from class to find the ATT of sex on the log wage rate.

2.c Discuss Results

(c) Show results:

(i) Both estimators (4) and (5),

(ii) With and without cross-fitting,

(iii) Using different first-step estimators for the propensity score $\hat{p}(x_i)$ and regression function $\hat{\mu}_0(x_i)$, including forests, neural networks, LASSO, and parametric models.

Discuss the results.

3 An Application

The file `data_for_HW4.csv` contains data from two independent sources, as indicated by the variable e. Both have data on the same outcome y, same treatment t, and the same set of pre-treatment variables $x.1, x.2, x.3, x.4, x.5$. The treatment in the first data source may have been targeted based on some or all of the x variables. The second data source is a fully randomized experiment. Both obey our other assumptions (SUTVA, consistency, CIA, overlap).

3.a Ignoring x Variables

Ignore the x variables to compute the ATE and a confidence interval for it in each of the data sources. Comment on your findings and possible explanations for them.

3.b Linear Model with Interactions

Use a linear model with interactions to obtain the CATEs in each data source, plot the distribution of the CATEs, obtain the ATE and its confidence interval. Compare your findings on the ATEs to the previous part.

3.c Doubly Robust Estimation

Combine the estimators of $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$ with a parametric logistic regression estimate of the propensity score $p(x) = \mathbb{P}[T = 1 \mid X = x]$ to estimate the ATE and confidence interval in each data source using the doubly robust estimator. Compare your findings on the ATEs to the previous two parts.

3.d Flexible/Nonparametric Versions

Replace your estimates of $\mu_t(x)$ and $p(x)$ with flexible/nonparametric/ML versions, and repeat the doubly robust estimation and inference. Try a few different nonparametric estimators for practice.

3.e Combined Model for Both Datasets

Propose and estimate a model (parametric or not) that combines and uses the two datasets as one. In other words, your model should have a single loss function, shared or common parameters, and appropriate assumptions as you deem fit. You must use data from both sources. Discuss your choice of specification and the properties of your proposed estimator.