Homework Assignment 1

Econ 31380 Causal Machine Learning
Max H. Farrell

Due October 18. Submit your answers on Canvas typed using LaTeXor Markdown.

1 Identification of Variance in Random Experiments

In class we discussed how the assumptions of randomized treatment assignment, SUTVA, and consistency were sufficient to identify average treatment effects such as $\tau = \mathbb{E}[Y(1) - Y(0)]$ and conditional averages like $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$.

(a) Prove that the variance of the individual treatment effects is not identified. That is, show that even though $\mathbb{E}[Y(1) - Y(0)]$ is identified, $\mathbb{V}[Y(1) - Y(0)]$ is not. Intuitively explain the source of the identification problem.

(b) In contrast, prove that the variance of the conditional average treatment effect, i.e., $\mathbb{V}[\tau(X)]$ is identified.

(c) Explain in a real world context why a decision would want to know $\mathbb{V}[Y(1) - Y(0)]$ and separately $\mathbb{V}[\tau(X)]$. For each $\mathbb{V}[Y(1) - Y(0)]$ and $\mathbb{V}[\tau(X)]$, explain how you would use this variance in research and in decision making.

2 Linear Regression in Randomized Experiments

Assume that $(y_i, x_i, t_i,), i = 1, \ldots, n$ is an iid sample from $(Y, X, T) \in \mathbb{R}^2 \times \{0, 1\}$ (having a vector of covariates changes nothing but notation). Further assume this is an ideal randomized experiment in the sense that (i) T is randomized such that $\mathbb{P}[T = 1 \mid X = x] = p$, which is bounded inside $(0, 1)$, (ii) x_i is realized prior to randomization, (iii) SUTVA and consistency hold.

First consider only Y and T.

(a) Using potential outcomes notation and specific assumptions to prove that (i) without loss of generality we can write $Y = Y_1 T + Y_0(1 - T) = \alpha + \beta T + \varepsilon$, (ii) the average treatment effect obeys $\tau := \mathbb{E}[Y_1 - Y_0] = \beta$, and (iii) the residuals ε are mean zero given the regressor.

(b) Show that the variance of ε is heteroskedastic in general. Under what conditions will it be homoskedastic?

(c) Taking this model to data, we obtain

$$(\widehat{\alpha}, \widehat{\tau}) = \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - bt_i)^2 . \tag{1}$$

Derive closed form solutions for the vector $\widehat{\theta} = (\widehat{\alpha}, \widehat{\tau})'$ using matrix notation and for the individual components using scalar notation. Give conditions for the solutions to exist and be unique.

(d) Prove that $\widehat{\tau} = \bar{Y}_1 - \bar{Y}_0$ where $\bar{Y}_1 = \sum_{i=1}^{n} y_i t_i / n_1$ with $n_1 = \sum_{i=1}^{n} t_i$, and similar for $\bar{Y}_0$.

(e) Use least squares algebra to derive the variance of $\widehat{\theta}$ conditional on $t_1, \ldots, t_n$ and find its probability limit. Give a consistent estimator. Is your estimator unbiased?

Now we will estimate the average treatment effect (ATE) using the plug-in principle. The ATE is $\tau = \mathbb{E}[Y(1) - Y(0)]$. The plug-in principle replaces unknown quantities with sample analogues (put hats on stuff) and replaces population averages with sample averages. Therefore we will consider

$$\widehat{\tau}_{\text{pi}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\tau}_i, \qquad \text{for some estimate } \widehat{\tau}_i = \widehat{Y_i(1) - Y_i(0)}.$$

(f) First, we run least squares on demeaned covariates separately in each group to obtain $\widehat{\theta}_t = (\widehat{\alpha}_t, \widehat{\beta}_t)'$ as

$$(\widehat{\alpha}_t, \widehat{\beta}_t) = \arg\min_{a,b} \sum_{i=1}^{n} \mathbf{1}\{t_i = t\} \left(y_i - a - b(x_i - \bar{x}_t)\right)^2. \tag{2}$$

Give closed form solutions for the vector $\widehat{\theta}_t = (\widehat{\alpha}_t, \widehat{\beta}_t)'$ using matrix notation and give conditions for the solutions to exist and be unique.

(g) Given $\widehat{\theta}_0$ and $\widehat{\theta}_1$, form predicted values for each potential outcome, i.e. give expressions for $\widehat{Y}_i(1)$ and $\widehat{Y}_i(0)$. Combine these counterfactual predictions with each observation's factual (observed) outcome to obtain estimates of the individual causal effect $\tau_i = Y_i(1) - Y_i(0)$. Denote these predicted values as $\widehat{\tau}_i$.

(h) Show that $\widehat{\tau}_{\text{pi}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\tau}_i = \widehat{\alpha}_1 - \widehat{\alpha}_0$, the latter being the intercepts from (2).

(i) Show how to obtain $\widehat{\tau}_{\text{pi}}$ using a single least squares regression and, if need be, taking linear combinations of its coefficients.

(j) Prove that $\widehat{\tau}_{\text{pi}}$ converges in probability to the average treatment effect.

(k) Carefully derive the asymptotic distribution of $\widehat{\tau}_{\text{pi}}$ and provide an estimator of the asymptotic variance. State your assumptions carefully and completely.

(l) Assume that, for $t = \{0, 1\}$, $Y(t) = \alpha_t + \beta_t X + \varepsilon_t$, with $\mathbb{E}[\varepsilon_t \mid X] = 0$ and $\mathbb{E}[\varepsilon_t^2 \mid X] = \sigma_{tX}^2$ is constant. Prove that the asymptotic variance from part **(k)** is weakly smaller than that of **(e)**, i.e. prove that $\widehat{\tau}_{\text{pi}}$ is at least as efficient as the difference in means. Under what conditions is $\widehat{\tau}_{\text{pi}}$ strictly more efficient?

(Here we prove an efficiency gain assuming linearity and homoskedasticity. The results holds much more generally as shown first by Lin (2013, Annals of Applied Statistics).)

3 Multiple Treatments

Consider studying two different binary treatments, $S \in \{0, 1\}$ and $T \in \{0, 1\}$. Assume we have access to experimental data where both S and T have been randomly assigned with constant (possibly different) probabilities, and an outcome Y and a pre-treatment covariate X are measured. The sample data are thus $(t_i, s_i, y_i, x_i), i = 1, \ldots, n$.

(a) Carefully define the potential outcomes in this setting. Make any assumptions explicit.

(b) Write down a linear regression model to recover the average of each potential outcome.

(c) Use estimates of the parameters in that model to estimate the average effect of treatment T. That is, for an individual drawn at random from the superpopulation, what is the estimated expected difference in their outcome under $T = 1$ versus $T = 0$?

(d) Using Monte Carlo simulations, compare this estimator the simple difference in means estimator for treatment T. Compare in terms of bias, consistency, and efficiency. Explain what you find.

4 Linear Models and Interactions

We are studying a marketing campaign for a website. We have data from 200 different markets (think of billboards advertising the website placed in different cities) and we observe:

- `visitors` = Total visitors to the website (in hundreds of thousands) coming from that market,
- `spend` = Total spent (in thousands of dollars) on advertising in that market.

This data is on Canvas in the file `marketing_data.csv`. Assume throughout that changes in spending *cause* changes in visitor numbers. The question is whether the spending on advertising is helping or hurting website traffic and by how much.

We start with the following (heteroskedastic) regression model with independent observations:

$$\texttt{visitors}_i = \beta_0 + \beta_T \times \texttt{spend}_i + \varepsilon_i, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_i^2), \qquad \text{correlation}(\texttt{spend}_\texttt{i}, \varepsilon_i) = 0.$$

(a) Show a scatter plot of the data along with the fitted regression line from this model. Give a precise & numerical causal interpretation of the results of running this linear regression How does spending cause visitors to change? Include a discussion of the statistical significance.

(b) Using the results above and your answer in **(a)**, if the goal is to increase the visitors to our website, should we *raise* or *lower* our spending?

Now we will break this analysis down by market type: the variable `city` is a binary variable taking values "`small`" and "`large`" according to city population size.

(c) Run a linear model of `visitors` on `spend` separately for each group separately. What do you find and how does this differ from what you found before? In your explanation, remember that spending *causes* visitors in this data.

(d) Show a scatter plot of the data with these two fitted regression lines, with the different market types shown in different colors.

(e) Write down a single regression model using `visitors`, `spend`, and `city`, that when taken to the data would yield exactly the four coefficient estimates in part **(c)** above. Match up your coefficients to the estimates in **(c)**.

(f) Give an expression for β_T, from the model shown before part **(a)**, in terms of the pieces of your model in part **(e)** and other features of the relationships between the three observed variables. Assume more things if you need to but be explicit. Use this expression to highlight what features of the data (i.e. the data generating process in the population) are leading to the differences between the separate fits in parts **(c)** and **(e)** compared to the combined fit in part **(a)**. Demonstrate these using the data.

(For future reference, this phenomenon is called "Simpsons Paradox".)