

Economics 696F, Causal Inference and Program Evaluation

Lecture Note 3: Unconfounded Treatment Assignment and Regression Analysis

1 Causal Model and Unconfoundedness

Next we turn to observational studies, where treatment assignment is not under our control.

Suppose $(T_i, Y_i(0), Y_i(1))$ are IID from some joint distribution, and as before, $Y_i := (1 - T_i)Y_i(0) + T_iY_i(1)$.

Last time, we showed that if treatment is randomly assigned (hence independent of the potential outcomes), the ATE is identified. However, in many economic studies, the treatment has not been randomly assigned. Are there other assumptions that would be sufficient to identify the treatment effect?

Suppose that in addition to the treatment and outcome, we also observe a background variable (or vector of variables) X_i , and we make the following assumption:

Unconfoundedness:

$$T_i \perp (Y_i(1), Y_i(0)) | X_i.$$

This says that T_i is independent of the potential outcomes *conditional* on X_i .

Unconfoundedness implies:

$$Pr(T_i = 1 | X_i = x, Y_i(1), Y_i(0)) = Pr(T_i = 1 | X_i = x).$$

Sometimes unconfoundedness is called the “selection on observables” assumption.

We also assume: for all x ,

$$0 < Pr(T_i = 1 | X_i = x) < 1,$$

This says that for each possible value of X_i , we have a positive probability of observing $T_i = 1$, and of observing $T_i = 0$.

Example: Consider a job training program.

T_i : indicator for receiving training

Y_i : employment indicator

X_i : binary variable equal to 1 if person i went to college, and equal to 0 if they only have a high school diploma.

Unconfoundedness implies:

$$Pr(T_i = 1 | X_i = 1, Y_i(1), Y_i(0)) = p_c$$

$$Pr(T_i = 1 | X_i = 0, Y_i(1), Y_i(0)) = p_{hs}$$

This says that for the subgroup with $X_i = 1$, we essentially have a randomized experiment with probability p_c of getting the treatment, and probability $1 - p_c$ of not getting the treatment. Likewise, for the subgroup with $X_i = 0$, we have a randomized experiment with probability p_{hs} of getting the treatment, and probability $1 - p_{hs}$ of not getting the treatment.

The second assumption implies that $0 < p_c < 1$ and $0 < p_{hs} < 1$, so that for each of the two groups, we don't observe all treated or all controls.

Example: suppose treatment effect is constant:

$$Y_i(1) - Y_i(0) = \tau, \quad \forall i$$

Then we can write:

$$Y_i = \alpha + \tau T_i + \epsilon_i,$$

where $\alpha := E[Y_i(0)]$, and $\epsilon_i := Y_i(0) - E[Y_i(0)]$. Then unconfoundedness is equivalent to the assumption that T_i is independent of ϵ_i .

Obviously, unconfoundedness is a very strong assumption and may not be appropriate in many circumstances. Whether or not it is reasonable will depend a great deal on the nature of X_i . It may be appropriate if X_i contains all the variables that we think are important for the decision to get the treatment.

Note that in general, $p_c \neq p_{hs}$, so that the treated and control groups could differ in their ratios of high school to college graduates. So just comparing treated and untreated groups (as we did in the randomized treatment case) will not be appropriate.

2 Identification under Unconfoundedness:

For now, focus on $ATE(x)$ and ATE . (Similar argument works for TT .)

Notice that

$$\begin{aligned} E[Y_i|T_i = 1, X_i = x] &= E[Y_i(1)|T_i = 1, X_i = x] \\ &= E[Y_i(1)|X_i = x] \end{aligned}$$

where the last line follows from the unconfoundedness assumption. Since $Pr(T_i = 1|X_i = x) > 0$ by assumption, we can consistently estimate $E[Y_i|T_i = 1, X_i = x]$, and therefore we can identify $E[Y_i(1)|X_i = x]$.

Likewise

$$\begin{aligned} E[Y_i|T_i = 0, X_i = x] &= E[Y_i(0)|T_i = 0, X_i = x] \\ &= E[Y_i(0)|X_i = x] \end{aligned}$$

so we can estimate $E[Y_i(0)|X_i = x]$ as well. Thus, we can identify $ATE(x)$.

Notice also that

$$ATE = E[ATE(X_i)] = \int ATE(x)dF_X(x),$$

where F_X is the (marginal) distribution of X_i .

If we have an estimate of $ATE(x)$, then we can average over the distribution of X_i to get an estimate of ATE . So, we have shown that ATE is identified under the unconfoundedness assumption.

Note: we have not made parametric assumptions about the joint distribution of (Y, T, X) . In that sense this is a nonparametric identification result. On the other hand, the unconfoundedness assumption is very strong and rules out many interesting treatment selection processes.

3 Estimation: Parametric Methods

We showed that

$$\begin{aligned} ATE(x) &= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \\ &= E[Y_i|T_i = 1, X_i = x] - E[Y_i|T_i = 0, X_i = x] \end{aligned}$$

So the key is to estimate the regression function $E[Y_i|T_i, X_i]$.

Suppose $X_i = x$ is a discrete variable. Consider the estimator:

$$\hat{\tau}(x) = \frac{\sum_i T_i 1(X_i = x) Y_i}{\sum_i T_i 1(X_i = x)} - \frac{\sum_i (1 - T_i) 1(X_i = x) Y_i}{\sum_i (1 - T_i) 1(X_i = x)}.$$

We are simply taking the treatment and control averages for the subsample with X_i equal to a particular value.

Then take sample analog to the equation $ATE = E[ATE(x)]$:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \hat{\tau}(X_i).$$

A nice feature of this estimator is that we avoid making strong assumptions on the form of $E[Y_i|T_i, X_i]$. However, if X_i takes on many values, there will be relatively few observations with any particular value of X_i , leading to high variance for $\hat{\tau}(x)$.

If X_i has a continuous distribution, then this estimator cannot be used. Under continuity, the probability of observing more than one sample unit with the same value of X_i is zero. In practice, we will have very few observations with the same value of X_i .

One possibility is to make some assumptions about the conditional regression function $E[Y|T, X]$. For example suppose we assume that

$$E[Y_i|T_i, X_i] = \beta_1 + T_i\beta_2 + X_i'\beta_3 + (T_i \cdot X_i)'\beta_4.$$

This implies:

$$E[Y_i|T_i = 0, X_i] = \beta_1 + X_i'\beta_3,$$

$$\begin{aligned} E[Y_i|T_i = 1, X_i] &= (\beta_1 + \beta_2) + X_i'(\beta_3 + \beta_4) \\ &= \gamma_1 + X_i'\gamma_2. \end{aligned}$$

So the regression line for the $T_i = 0$ subgroup could have a different slope and intercept from the regression line for the $T_i = 1$ subgroup.

(We could include transformations of X_i (such as powers of X_i) as well, and get a fairly general regression specification.)

We could then estimate this regression function by OLS, and then estimate $\tau(x)$ by:

$$\begin{aligned} \hat{\tau}(x) &= E[Y|T = 1, X = x] - E[Y|T = 0, X = x] \\ &= \hat{\beta}_1 + \hat{\beta}_2 + x'(\hat{\beta}_3 + \hat{\beta}_4) - \hat{\beta}_1 - x'\hat{\beta}_3 \\ &= \hat{\beta}_2 + x'\hat{\beta}_4. \end{aligned}$$

Then the estimate of the overall average treatment effect would be

$$\begin{aligned} \hat{\tau} &= \frac{1}{n} \sum_{i=1}^n [\hat{\beta}_2 + X_i'\hat{\beta}_4] \\ &= \hat{\beta}_2 + \frac{1}{n} \sum_{i=1}^n X_i'\hat{\beta}_4 \\ &= \hat{\beta}_2 + \bar{X}'\hat{\beta}_4. \end{aligned}$$

We can get an even nicer expression for $\hat{\tau}$ if we run the regression with $(X_i - \bar{X})$ in place of X_i . Then we will have

$$\hat{\tau} = \hat{\beta}_2.$$

Notice that the unconfoundedness assumption provides a link between conventional regression parameters and the causal parameter τ . Thus, it is possible to interpret regression parameters as causal, but only under somewhat strong assumptions about the selection into treatment and control groups.

4 Estimation via Nonparametric Regression

Both the estimators for ATE that we considered previously were of the form:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \hat{m}_1(X_i) - \hat{m}_0(X_i),$$

where $\hat{m}_1(x)$ is an estimate of

$$m_1(x) := E[Y_i|T_i = 1, X_i = x]$$

and $\hat{m}_0(x)$ is an estimate of

$$m_0(x) := E[Y_i | T_i = 0, X_i = x].$$

In the case of a continuous covariate, we assumed that m_0 and m_1 were both linear functions. This is clearly a strong assumption, so in this note we'll see how much we can relax it.

4.1 Brief Introduction to Nonparametric Regression

Let's start by considering a simpler problem. Suppose that for $i = 1, \dots, n$, (Y_i, X_i) are IID from some joint distribution. We assume Y_i and X_i are scalar and continuously distributed, and that X_i has compact support in \mathbb{R} . We want to estimate the function

$$m(x) := E[Y_i | X_i = x]$$

without making strong parametric assumptions. We will review some standard methods, omitting technical details for the time being.

Series estimator

The basic idea behind series estimators is to approximate the regression function $m(x)$ by a series

$$\begin{aligned} m(x) &\approx \theta_0 g_0(x) + \theta_1 g_1(x) + \dots + \theta_J g_J(x) \\ &= \sum_{j=0}^J \theta_j g_j(x). \end{aligned}$$

Here, the $g_j(\cdot)$ are called *basis functions*, and are chosen in advance, hopefully in such a way that linear combinations of them lead to a wide range of functional forms. We then use some method to estimate the weights θ_j using the observed data.

The simplest example is to take the polynomial functions:

$$\begin{aligned} g_0(x) &= 1 \\ g_1(x) &= x \\ g_2(x) &= x^2 \\ &\vdots \\ g_J(x) &= x^J \end{aligned}$$

We could then estimate the θ_j by least squares:

$$\min_{\theta_0, \dots, \theta_J} \sum_{i=1}^n \left(Y_i - \sum_{j=0}^J \theta_j X_i^j \right)^2 = \min_{\theta_0, \dots, \theta_J} \sum_{i=1}^n (Y_i - \theta_0 - \theta_1 X_i - \dots - \theta_J X_i^J)^2.$$

This amounts to a polynomial regression of Y on X .

A justification for this is the following famous theorem:

Theorem 1 (Stone-Weierstrass) Let $m(\cdot)$ be a continuous bounded function with compact domain $\mathcal{X} \subset \mathbb{R}$. Then, for any $\epsilon > 0$, there exists a polynomial function

$$l(x) = \sum_{j=0}^{\infty} \theta_j x^j,$$

such that

$$\sup_{x \in \mathcal{X}} |m(x) - l(x)| < \epsilon.$$

This says that any continuous function can be approximated arbitrarily well by some polynomial, and suggests that if J is reasonably large, we can approximate $m(x) = E[y_i | x_i = x]$ by a J th order polynomial.

How do we choose J ? Clearly, we cannot have $J \geq n$, because then we will have more parameters $\theta_0, \dots, \theta_J$ to estimate than we have data points. We want to have $(J + 1)$ smaller than n , but if we allow J to grow as $n \rightarrow \infty$, we can still have $\hat{m}(x) \xrightarrow{p} m(x)$ at every x .

Practically speaking, there tends to be a couple of different approaches taken to choosing J . One is to keep adding terms until the last term is no longer statistically significant (using the standard t-statistic).

An approach that seems to work well in practice is v -fold cross validation. The idea is to try to estimate the mean squared error $E[(Y_i - \hat{m}(X_i))^2]$, and choose J to minimize the estimated MSE.

Divide the sample into v equally sized groups. For each i , let $\hat{m}_J^{-i}(\cdot)$ be the estimated regression function using observations from the $v - 1$ groups that don't include observation i , and using a J th order polynomial specification. Then form

$$\widehat{MSE}(J) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{m}_J^{-i}(X_i))^2.$$

We then choose the order J which minimizes this approximate MSE.

There are other sets of basis functions that can be used, besides the ordinary polynomials. Fourier analysis typically uses sine and cosine-based functions. So-called neural network methods use a different set of basis functions related to the logistic CDF.

Series Logit Estimator

Suppose that Y_i is binary instead of being continuous, so that

$$E[Y_i | X_i = x] = Pr(Y_i = 1 | X_i = x).$$

In principle, we could continue to use the series estimator without modification. However, this has the undesirable property that it can lead to estimates $\hat{m}(x)$ outside the unit interval $[0, 1]$. This is the usual problem with the linear probability model.

One possibility is to use a transformation to keep the predicted probabilities between 0 and 1. For example, we can use the inverse-logit transform:

$$\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)},$$

and approximate the conditional probability function as:

$$\begin{aligned} Pr(Y_i = 1|X_i = x) &\approx \Lambda(\theta_0 + \theta_1 x + \dots + \theta_J x^J) \\ &= \frac{\exp(\theta_0 + \theta_1 x + \dots + \theta_J x^J)}{1 + \exp(\theta_0 + \theta_1 x + \dots + \theta_J x^J)}. \end{aligned}$$

Notice that Λ is strictly monotone and invertible, so for any function $m(x)$, we can define

$$g(x) := \Lambda^{-1}(m(x)),$$

so that

$$m(x) = \Lambda(g(x)).$$

This suggests that by choosing a sufficiently large J and appropriate coefficients $\theta_0, \dots, \theta_J$, we can approximate any conditional probability function $Pr(Y_i = 1|X_i = x)$ very well.

To estimate the $\theta_0, \dots, \theta_J$, one natural way is to use maximum likelihood:

$$(\hat{\theta}_0, \dots, \hat{\theta}_J) = \arg \max_{\theta_0, \dots, \theta_J} \prod_{i=1}^n [\Lambda(\theta_0 + \theta_1 x + \dots + \theta_J x^J)]^{Y_i} \times [1 - \Lambda(\theta_0 + \theta_1 x + \dots + \theta_J x^J)]^{1-Y_i}$$

Nearest Neighbor estimator

If X_i were discrete, then a natural estimator for $E[Y_i|X_i = x]$ would be

$$\hat{m}(x) = \frac{\sum_i 1(X_i = x)Y_i}{\sum_i 1(X_i = x)}.$$

That is, just take the observations with $X_i = x$ and average their outcome values. The k -nearest neighbor estimator is defined as

$$\hat{m}(x) := \frac{1}{k} \sum_{i: X_i \in N_k(x)} Y_i,$$

where $N_k(x)$ is set of the k values of X_i that are closest to x . So we find the k observations with X values closest to x , and average their outcomes. The idea is that if $m(x)$ is relatively smooth, so it does not change too much as x varies in a small neighborhood, then taking an average over values close to x should give a reasonable approximation.

We can calculate this at a set of different values of x , producing an estimated regression function that is easily plotted.

It's useful to think about what happens for different values of k .

If $k = n$, then we are using all of the observations, and $\hat{m}(x)$ just becomes the sample average of Y_i . This produces a perfectly flat estimated function. On the one hand, we are using a lot of data to do the averaging, so the variance is relatively low. But, if $m(x)$ is not actually constant, then our estimator will be biased for many particular values of x .

At the other extreme, we could set $k = 1$. This uses just one observation in the sample average. Since we are only using observations with X_i very close to x , the bias should be relatively small, but at the cost of using very few observations - hence high variance.

How to pick k ? We can use the same v -fold cross validation technique as in the series case. Now, instead of varying the number of terms J , we try different values of k and pick the one that minimizes the CV estimate of MSE.

Kernel estimator

The k -nearest-neighbor method takes averages in “neighborhoods” $N_k(x)$ of a point x , where the neighborhoods are defined in such a way as to contain a fixed number k of observations. An alternative approach is to define a neighborhood as a fixed interval about x . For example, we could take intervals of the form $[x - c, x + c]$, where c is some constant we choose, leading to an estimator:

$$\hat{m}(x) := \frac{\sum_{i=1}^n 1(X_i \in [x - c, x + c])Y_i}{\sum_{i=1}^n 1(X_i \in [x - c, x + c])}.$$

As in the nearest neighbor approach, the intuition is to take local averages of outcomes Y_i with X_i close to x .

We can extend this idea, by weighting the observations by how close they are to x . Let $K(u)$ be a function that is continuous, symmetric about 0 and bounded, and satisfying:

$$\int K(u)du = 1.$$

We call K a kernel. An example is $K(u) = \phi(u)$, the standard normal density function. We weight observations by

$$\frac{K\left(\frac{X_i - x}{h}\right)}{h},$$

where h is called a bandwidth. Note that as $h \rightarrow 0$, only observations with X_i very close to x will receive any weight.

The local averaging estimator then becomes:

$$\hat{m}(x) := \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}.$$

This is called the kernel regression estimator, or the Nadaraya-Watson estimator.

The estimator depends on the choice of kernel K , and on the bandwidth h . The most popular choice for the kernel function is the Epanechnikov kernel:

$$K(u) = 0.75(1 - u^2)1(|u| \leq 1),$$

which has some optimality properties and is easy to calculate. In practice, the Epanechnikov kernel and the normal kernel usually give similar results.

The choice of bandwidth makes a big difference in the resulting estimate. Choosing h small leads the estimator to only use observations very close to x , leading to a “wiggly” function. If we use larger h , the estimated function will be smoother, but since we are averaging observations with quite different X_i values, we expect more bias.

As before, we can use v -fold CV. There are a number of other methods for choosing the bandwidth.

4.2 Nonparametric Regression for Treatment Effects

Nonparametric Regression for ATE

Return to our treatment effects setting, and recall:

$$\begin{aligned} ATE(x) &:= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \\ &= E[Y_i|T_i = 1, X_i = x] - E[Y_i|T_i = 0, X_i = x] \\ &= m_1(x) - m_0(x) \end{aligned}$$

Also recall:

$$ATE = E[ATE(X_i)] = E[m_1(X_i) - m_0(X_i)].$$

We can use the data to estimate m_1 and m_0 by \hat{m}_1 and \hat{m}_0 using any of the techniques described above.

For example, we can take the observations with $T_i = 1$, and regress Y_i on $1, X_i, X_i^2, \dots, X_i^J$ to obtain an estimate $\hat{m}_1(x)$, and likewise with the $T_i = 0$ observations to get $\hat{m}_0(x)$. Then form:

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \hat{m}_1(X_i) - \hat{m}_0(X_i).$$

Note that we are taking each observed value of X_i , and evaluating \hat{m}_1 and \hat{m}_0 at each such value, then averaging.

Nonparametric Regression for TT

Recall

$$\begin{aligned} TT &:= E[Y_i(1) - Y_i(0)|T_i = 1] \\ &= E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1]. \end{aligned}$$

Note that $E[Y_i(1)|T_i = 1] = E[Y_i|T_i = 1]$, so the first term can be estimated simply by taking the sample average of the treated observations:

$$E[\widehat{Y_i(1)}|T_i = 1] = \frac{\sum_{i=1}^n 1(T_i = 1)Y_i}{1(T_i = 1)}.$$

How to estimate the second term? By the unconfoundedness assumption,

$$E[Y_i(0)|T_i = 1, X_i = x] = E[Y_i(0)|T_i = 0, X_i = x] = E[Y_i|T_i = 0, X_i = x].$$

And by iterated expectations

$$E[Y_i(0)|T_i = 1] = E\left[E[Y_i(0)|T_i = 1, X_i] \middle| T_i = 1\right].$$

So we can use the estimate

$$E[\widehat{Y_i(0)}|T_i = 1] = \frac{\sum_{i=1}^n 1(T_i = 1)\hat{m}_0(X_i)}{\sum_{i=1}^n 1(T_i = 1)}.$$

This leads to the following estimator of TT:

$$\hat{\tau}_t := \frac{\sum_{i=1}^n 1(T_i = 1)[Y_i - \hat{m}_0(X_i)]}{\sum_{i=1}^n 1(T_i = 1)}.$$

For example, suppose we use 1-nearest neighbor matching for \hat{m}_0 . Basically, what we are doing is taking each treated outcome, and finding a control unit that has the closest value of X . We subtract the treated outcome from its “matched” control outcome, and average. This is a type of matching estimator, which are studied in greater detail in Abadie and Imbens (2004)

5 Additional References:

Here are some useful books on nonparametric regression techniques:

Hastie, T, Tibshirani, R, and Friedman, J., 2001 *The Elements of Statistical Learning*, New York: Springer.

Efromovich, S., 1999, *Nonparametric Curve Estimation*, New York: Springer.