# Treatment Effects

Whitney K. Newey
MIT

March 2007

The treatment effects literature is about how some outcome of interest, such as earnings, is affected by some treatment, such as a job training program. Evidently such treatment effects must be related to structural models, where the outcome of interest is the left hand side variable and the treatment is a right-hand side variable. Indeed, as we will see, treatment effects can be thought of as coming from a linear structural model with random coefficients. Treatment effect models do have a terminology and set up all their own though, so to help understand the literature it is important to set them up the way others do.

To do so, let $i$ index individuals and $D_i$ denote a treatment indicator, equal to 1 if a person is treated, and equal to 0 otherwise. For example, $D_i = 1$ might correspond to enrollment in some training program or to some medical treatment. To describe the treatment effect, we need to define two other variables. Let $Y_{i0}$ denote the potential outcome that would occur when person $i$ is not treated $(D_i = 0)$ and $Y_{i1}$ the potential outcome when they are treated $(D_i = 1)$. Clearly these are not both observed. One of them will be "counterfactual", an outcome that would have occurred if a different treatment had been given. The observed outcome will be

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}.$$

The treatment effect for individual $i$ is given by

$$\beta_i = Y_{i1} - Y_{i0}.$$

This object is clearly not identified, because only one of the potential outcomes is observed. There are several objects that may be interesting that are identified under certain conditions. One of these is the average treatment effect, given by

$$ATE \overset{def}{=} E[\beta_i].$$

This describes the average over the entire population of the individual treatment effects. Another interesting object is the average effect of treatment on the treated, given by

$$ATT \overset{def}{=} E[\beta_i | D_i = 1].$$

This gives the average over the subpopulation of treated people of the treatment effect. A third important object that is also of interest in the literature is called the local average treatment effect. It will be described below.

To help understand the treatment framework and the various effects, it helps to relate this to a regression model with random coefficients. By the equation for $Y_i$ given above,

$$\begin{aligned} Y_i &= Y_{i0} + (Y_{i1} - Y_{i0})D_i = \alpha_i + \beta_i D_i, \\ \alpha_i &= Y_{i0}, \beta_i = Y_{i1} - Y_{i0}. \end{aligned}$$

Thus we see that $Y_i$ follows a linear model where the treatment effect $\beta_i$ is the coefficient of $D_i$ and the constant $\alpha_i$ and slope $\beta_i$ may vary over individuals. The ATE is then the average of the slope over the entire population and the ATT is the average of the slope over the subset of the population where $D_i = 1$.

This random coefficient set up also helps place the treatment effects environment in a proper historical context. The coefficient $\beta_i = Y_{i1} - Y_{i0}$ is sometimes called a "counterfactual" because it describes how $Y_i$ would have been different if $D_i$ had been different. In the context of demand and supply systems we are familiar with such objects as "movements along a curve." This kind of object was considered in economics as early as Wright (1928), who gives a nice explanation of "movements along a curve" in a supply and demand setting. Similarly, the average treatment effect is just the expected value of the random coefficient in a linear model, i.e. the average slope of the curve.

The ATE and ATT will be identified and can be estimated under various assumption. Here we will discuss various cases in which these objects are identified. The proofs of identification will consist of showing how the objects can be written in terms of expectations of the data.

We begin with the simplest case.

## Constant Treatment Effects

A simple special case of this model is constant treatment effects where $\beta_i = \bar{\beta}$, i.e. where the treatment effect is constant across individuals. Here the ATE and ATT is simply $\bar{\beta}$. In this case, for $\bar{\alpha} = E[\alpha_i]$ and $\varepsilon_i = \alpha_i - \bar{\alpha}$,

$$Y_i = \bar{\alpha} + \bar{\beta} D_i + \varepsilon_i.$$

Here the model reduces to a simple linear model with an additive disturbance and constant coefficients. In contrast, the general model is also a linear model with additive disturbance but random slope coefficient. Note here the equivalence between having a random $\alpha_i$ and having a constant plus disturbance $\alpha_i = \bar{\alpha} + \varepsilon_i$.

We can identify and estimate $\bar{\beta}$ and $\bar{\alpha}$ in the usual way if we have an instrument $Z_i$ that is uncorrelated with $\varepsilon_i$ and correlated with $D_i$, that is

$$\begin{aligned} 0 &= Cov(Z_i, \varepsilon_i) = Cov(Z_i, \alpha_i) = Cov(Z_i, Y_{i0}), \\ Cov(Z_i, D_i) &\neq 0. \end{aligned}$$

In this case the coefficient is identified from the usual IV equation

$$\bar{\beta} = Cov(Z_i, Y_i)/Cov(Z_i, D_i).$$

This coefficient can be estimated in the usual way by replacing population covariances by sample covariances. In summary, there is not much new here, except terminology of putting standard model with dummy endogenous variable in a treatment effects framework, as ”constant treatment effect.”

Constant treatment effects is too strong for many settings. It would say that effect of training on earnings or of smaller class size on education is the same for every individual. This seems unlikely to hold in practice. Instead we would like to allow $\beta_i$ to vary over individuals.

## Random Assignment

Random assignment means that whether or not a person is treated does not depend on their outcomes. The specific statistical assumption that we make is that

$$E[Y_{i0}|D_i] = E[Y_{i0}],$$

i.e. that the mean of the nontreated variable does not depend on treatment status. Equivalently we can say that $E[\alpha_i|D_i] = 0$. This is slightly more general than independence, because it allows the higher-order moments of $Y_{i0}$ to depend on $D_i$. However, it seems difficult to think of environments where the mean assumption would be true without full independence.

To see what happens under this assumption note first that

$$E[\beta_i|D_i]D_i = \left\{ \begin{array}{c} 0, \quad D_i = 0, \\ E[\beta_i|D_i = 1], D_i = 1 \end{array} \right. = E[\beta_i|D_i = 1]D_i.$$

Then under the mean independence assumption,

$$\begin{aligned} E[Y_i|D_i] &= E[\alpha_i + \beta_i D_i|D_i] = E[\alpha_i] + E[\beta_i|D_i]D_i \\ &= E[\alpha_i] + E[\beta_i|D_i = 1]D_i. \end{aligned}$$

Here the dummy variable regression of $Y_i$ on a constant $D_i$ has its slope coefficient the ATT. If in addition we assume that the mean of $Y_{i1}$ does not depend on $D_i$, i.e. if we assume that

$$E[Y_{i1}|D_i] = E[Y_{i1}],$$

Then we find that $ATE = ATT$, since

$$\begin{aligned} E[\beta_i|D_i &= 1] = E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 1] \\ &= E[Y_{i1}] - E[Y_{i0}] = E[\beta_i]. \end{aligned}$$

Summarizing, we find that when $Y_{i0}$ is mean independent of $D_i$ that the $ATT$ is identified as the dummy coefficient in a regression of the outcome variable $Y_i$ on a constant and the treatment dummy variable. We also find that if, in addition, $Y_{i1}$ is mean independent of $D_i$ then the $ATE$ is also this coefficient. Of course, this coefficient can be estimated by a linear regression of $Y_i$ on $(1, D_i)$. Further, as always, that linear regression coefficient is just the difference of means of $Y_i$ for the treated and untreated observations.

## Discussion

Random assignment is too strong for many applications. Often individuals can choose whether to accept the treatment or not, e.g. by dropping out of the sample if they don't like the treatment conditions. They can opt out of training programs, or not take medical treatment. If these decisions are related to $(\alpha_i, \beta_i)$ then we do not have independence of $(\alpha_i, \beta_i)$ and $D_i$. In terms of the linear model $Y_i = \alpha_i + \beta_i D_i$ we have possible endogeneity, where $D_i$ may be correlated with the random coefficients $\alpha_i$ and $\beta_i$. This is a more severe problem than the usual case because the slope $\beta_i$ also may be correlated with $D_i$.

There are two approaches to this problem. One (familiar) one is instrumental variables (IV). The second approach is called "selection on observables." In that approach conditioning on some observable variables removes the correlation between $D_i$ and $(\alpha_i, \beta_i)$. Because IV is a most familiar and common approach we will first consider IV.

## IV Identification of Treatment Effects

In the usual linear model, of which the constant treatment effects is a special case, the assumptions that are needed for the identification of the slope is that the instrument is uncorrelated with the disturbance and correlated with $D_i$. Similar conditions will be used for IV identification of treatment effects. Let $Z_i$ be an instrument. We will assume throughout that

$$E[\alpha_i|Z_i] = E[Y_{i0}|Z_i] = E[Y_{i0}] = E[\alpha_i], .$$

i.e. that the outcome without treatment is mean independent of the instrument.

We also will focus on the case where $Z_i$ is also a dummy variable, i.e. where $Z_i \in \{0,1\}$, with $P = \Pr(Z_i = 1)$ and $0 < P < 1$. (Question: Why do we assume $0 < P < 1$ ?). For a dummy instrument there is a useful formula for the covariance between the instrument and any other random variable $W_i$. Specifically, we have

$$\begin{aligned} Cov(W_i, Z_i) &= E[W_i Z_i] - E[W_i]E[Z_i] = (\frac{E[W_i Z_i]}{P} - E[W_i])P \\ &= (E[W_i|Z_i = 1] - E[W_i])P \\ &= \{E[W_i|Z_i = 1] - (PE[W_i|Z_i = 1] + (1-P)E[W_i|Z_i = 0])\}P \\ &= (E[W_i|Z_i = 1] - E[W_i|Z_i = 0])P(1-P). \end{aligned}$$

That is, the covariance between $W_i$ and $Z_i$ is the difference of the conditional mean at the two values of $Z$ times $P(1-P)$.

This formula has two useful implications. The first is that mean independence of $Y_{i0}$ from $Z_i$ is equivalent to $Y_{i0}$ being uncorrelated with $Z_i$. This occurs since $Cov(Z_i, Y_{i0}) = 0$ if and only if $E[W_i|Z_i = 1] = E[W_i|Z_i = 0]$. A second useful implication is a formula for the limit of the IV estimator of the slope, given by

$$\frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

This is often referred to the Wald IV formula, referring to work where Wald suggested using dummy variables as an IV solution to the measurement error problem.

In general, under mean independence of $\alpha_i$ from $Z_i$, it does not seem like the IV formula identifies the ATT, the ATE, or anything useful. Plugging in $Y_i = \alpha_i + \beta_i D_i$, and using mean independence of $\alpha_i$ we find

$$\begin{aligned} \frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)} &= \frac{E[\alpha_i|Z_i = 1] - E[\alpha_i|Z_i = 0] + E[\beta_i D_i|Z_i = 1] - E[\beta_i D_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \\ &= \frac{E[\beta_i D_i|Z_i = 1] - E[\beta_i D_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}. \end{aligned}$$

In general, the problem is that $\beta_i$ and $D_i$ are correlated, so that (apparently) it is not possible separate them out in general. There are two interesting, specific cases though where something important is identified. They are random intention to treat and local average treatment effects.

## Random Intention to Treat

A common occurrence in medical trials is that people are randomly assigned to treatment but that not all take the treatment. Here $Z_i$ represents the assignment to treatment, with $Z_i = 1$ is individual $i$ is assigned to be treated and $Z_i = 0$ if they are not. In this setting, the only ones who are treated (i.e. for which $D_i = 1$) will be those who were randomly assigned to treated. It turns out that in this case IV gives the ATT. This finding, due to Imbens and Rubin, has led to the widespread use of IV in biostatistics.

To show that IV gives the ATT, note that $Z_i = 0$ will not be treated, i.e. $D_i = 0$ when $Z_i = 0$. Then

$$\frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)} = \frac{E[\beta_i D_i|Z_i = 1] - 0}{E[D_i|Z_i = 1] - 0} = \frac{E[\beta_i D_i|Z_i = 1]}{E[D_i|Z_i = 1]}.$$

Also, note that $D_i = 1$ implies $Z_i = 1$, so that $\{D_i = 1\} \subset \{Z_i = 1\}$. Therefore, $E[\beta_i|D_i = 1, Z_i = 1] = E[\beta_i|D_i = 1] = ATT$. Also, it follows similarly to the reasoning above that

$$D_i E[\beta_i|D_i, Z_i = 1] = D_i E[\beta_i|D_i = 1, Z_i = 1] = D_i \cdot ATT.$$

By iterated expectations it follows that

$$E[\beta_i D_i|Z_i = 1] = E[D_i E[\beta_i|D_i, Z_i = 1]|Z_i = 1] = ATT \cdot E[D_i|Z_i = 1].$$

Then dividing gives

$$\frac{Cov(Z_i, Y_i)}{Cov(Z_i, D_i)} = \frac{E[\beta_i D_i|Z_i = 1]}{E[D_i|Z_i = 1]} = \frac{ATT \cdot E[D_i|Z_i = 1]}{E[D_i|Z_i = 1]} = ATT.$$

## The Local Average Treatment Effect

A second case where an interesting treatment effect is identified by IV involves independence and monotonicity conditions. Consider the following conditions:

*Independence:* $D_i = \Pi(Z_i, V_i)$ *and* $(\beta_i, V_i)$ *is independent of* $Z_i$;

*Monotonicity:* $\Pi(1, V_i) \geq \Pi(0, V_i)$ *and* $\Pr(\Pi(1, V_i) > \Pi(0, V_i)) > 0$.

The independence condition says that there is a reduced form $\Pi(z, v)$ with a disturbance $V_i$ that may be a vector and enters nonlinearly. An example is a threshold crossing model where $D_i = 1(Z_i + V_i > 0)$. The monotonicity condition changing the instrument only moves the treatment one direction. This condition is satisfied in a threshold crossing model. The reduced form is sometimes called the selection equation, with a person being selected into treatment when $\Pi(z, v) = 1$.

Under these conditions it turns out that IV identifies an average of $\beta_i$ over a subpopulation that is referred to as the Local Average Treatment Effect (LATE). This effect is defined as

$$LATE = E[\beta_i|\Pi(1, V_i) > \Pi(0, V_i)].$$

This object is the average of the treatment effect over the individuals whose behavior would be different if the instrument were changed. This object may often be a parameter of interest. For example, in a model where $Y_i$ is the log of earnings, $D_i$ is completing high school, and $Z_i$ is a quarter of birth dummy, LATE is the average effect of a high school education over all those dropouts who would have remained in school had their quarter of birth been different and for those who remained in school but would have dropped out if their quarter of birth were different. Thus, IV estimates the average returns to completing high school for potential dropouts. This is an interesting parameter, although it is not the returns to schooling over the whole population.

To show that IV give LATE under independence and monotonicity, let $T_i = \Pi(1, V_i) - \Pi(0, V_i)$. Then we have

$$\begin{aligned} E[\beta_i D_i|Z_i &= 1] - E[\beta_i D_i|Z_i = 0] \\ &= E[\beta_i \Pi(1, V_i)|Z_i = 1] - E[\beta_i \Pi(0, V_i)|Z_i = 0] \\ &= E[\beta_i \Pi(1, V_i)] - E[\beta_i \Pi(0, V_i)] = E[\beta_i T_i]. \end{aligned}$$

It follows similarly that

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = E[T_i].$$

By monotonicity, $T_i$ is a dummy variable, taking the value zero or one. Therefore we have

$$\frac{\text{cov}(Z_i, Y_i)}{\text{cov}(Z_i, D_i)} = \frac{E[\beta_i T_i]}{E[T_i]} = E[\beta_i|T_i = 1] = E[\beta_i|\Pi(1, V_i) > \Pi(0, V_i)].$$

## LATE Empirical Example

An empirical example is provided by the Angrist and Krueger (1991) study of the returns to schooling using quarter of birth as an instrument. We consider data drawn from the 1980 U. S. Census for males born in 1930-1939, as in Donald and Newey (2001, "Choosing the Number of Instruments," Econometrica). The 2SLS estimator with 3 instruments is .1077 with standard error .0195 and the FULL estimator with 180 instruments is .1063 with standard error .0143 (corrected for many instruments). Thus we find returns to schooling of "potential dropouts" is about 11 percent.

## Selection on Observables

The other kind of model that has been used to identify treatment effects is one where conditioning on observable (or identifiable) variables $X_i$ makes the treatment behave as if it were randomly assigned. This is like removing endogeneity in a linear equation by adding regressors. The conditioning variables are like omitted regressors, which remove the endogeneity when they are included. The specific assumption that is made is

$$E[Y_{i0}|X_i, D_i] = E[Y_{i0}|X_i].$$

In word, it is assumed that $Y_{i0}$ is mean independent of $D_i$ conditional on $X_i$. This assumption is analogous to the previous one that $E[Y_{i0}|D_i] = E[Y_{i0}]$, being a conditional version of that hypothesis.

One concern with this kind of assumption is the source for the variables $X_i$. There are some economic models where such variables are implied by the model. However in many cases in applications these variables $X_i$ are chosen without reference to a model. In those cases identification is fragile, requiring specifying just the right $X_i$. Conditional mean independence that holds for $X_i$ need not hold for a subset of $X_i$ nor when additional variables are added to $X_i$.

This assumption allows identification of the ATT, with one additional condition. Let $\mathcal{X}$ denote the support of $X_i$ (the smallest closed set having probability one), and $\mathcal{X}_0$ and $\mathcal{X}_1$ the support of $X_i$ conditional on $D_i = 0$ and $D_i = 1$ respectively. The additional condition is the common support condition that

$$\mathcal{X} = \mathcal{X}_0 = \mathcal{X}_1.$$

This assumption is necessary and sufficient for $E[Y_i|X_i, D_i = 1]$ and $E[Y_i|X_i, D_i = 0]$ to be well defined for all $X_i$. It is verifiable and may or may not be satisfied in practice.

The common support condition and conditional mean independence give

$$\begin{aligned} E[Y_i|X_i, D_i &= 1] - E[Y_i|X_i, D_i = 0] = E[\alpha_i|X_i, D_i = 1] - E[\alpha_i|X_i, D_i = 0] + E[\beta_i|X_i, D_i = 1] \\ &= E[\beta_i|X_i, D_i = 1]. \end{aligned}$$

The object $E[\beta_i|X_i, D_i = 1]$ is a conditional version of the ATT. By iterated expectations the ATT is then identified as the expectation over $X_i$ of this difference given $D_i = 1$, that is

$$ATT = E[\beta_i|D_i = 1] = E[\{E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]\}|D_i = 1].$$

The ATE can also be obtained if we assume that $Y_{i1}$ is conditional mean independent of $D_i$ conditional on $X_i$. In that case

$$\begin{aligned} E[\beta_i|X_i, D_i &= 1] = E[Y_{i1}|X_i, D_i = 1] - E[Y_{i0}|X_i, D_i = 1] \\ &= E[Y_{i1}|X_i,] - E[Y_{i0}|X_i] = E[\beta_i|X_i]. \end{aligned}$$

Therefore

$$ATE = E[\beta_i] = E[\{E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]\}]$$

Unlike the unconditional case the ATE is a different function of the data distribution than the ATT. The ATE is obtained by averaging $E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]$ over all $X_i$ while the ATT is obtained by averaging over just $D_i = 1$.

Estimating the ATT and ATE under these conditional restrictions is a challenge. Notice that they depend on conditional expectations. Usually we will not want to assume that these conditional expectations have any particular functional form. Consequently, we will want to use nonparametric regression estimators, which will be discussed later in the course.

Nonparametric estimation is difficult when the dimension of $X_i$ is large. This is often referred to as the "curse of dimensionality." Some have tried to reduce the curse of dimensionality using the “propensity score” $P(X)$, which is defined as the conditional probability of being treated (or "selected") given $X$, i.e.

$$P(X_i) = \Pr(D_i = 1|X_i) = E[D_i|X_i].$$

It turns out that the conditional mean independence of $Y_{i0}$ given $X_i$ implies conditional mean independence given $P(X_i)$. Thus, if $P(X_i)$ were known, it would be possible to identify and estimate the ATE and ATT using a one dimensional conditioning variable rather than a multidimensional variable $X_i$. Specifically, if $E[Y_{i0}|X_i, D_i] = E[Y_{i0}|X_i]$ and $0 < P(X_i) < 1$ with probability one then $E[Y_{i0}|P(X_i), D_i] = E[Y_{i0}|P(X_i)]$, so that reasoning like that above gives

$$ATT = E[\{E[Y_i|P(X_i), D_i = 1] - E[Y_i|P(X_i), D_i = 0]\}|D_i = 1].$$

If, in addition, $E[Y_{i0}|X_i, D_i] = E[Y_{i0}|X_i]$ then

$$ATE = E[\{E[Y_i|P(X_i), D_i = 1] - E[Y_i|P(X_i), D_i = 0]\}].$$

Thus, ATE and ATT are expectations of nonparametric functions of two variables, $P(X_i)$, and $D_i$.

If $P(X_i)$ is completely unknown and unrestricted there is no known advantage for conditioning on the propensity score, since $P(X_i)$ is also a function of a high-dimensional argument. Thus, it appears that any advantage for using the propensity score will depend on knowing more about $P(X)$ than about $E[Y_i|X_i, D_i]$.

It remains to prove that independence conditional on $X$ implies independence conditional on $P(X_i)$. For notational simplicity let $P_i = P(X_i)$. We will prove the result for a general variable $W_i$. The general result will then apply to both $Y_{i0}$ and $Y_{i1}$. To prove that $E[W_i|X_i, D_i] = E[W_i|X_i]$ implies $E[W_i|P_i, D_i] = E[W_i|P_i]$, note that by iterated expectations, $E[D_i|P_i] = E[E[D_i|X_i]|P_i] = P_i$. By iterated expectations again,

$$\begin{aligned} E[W_i|P_i, D_i &= 1] = E[E[W_i|X_i, D_i = 1]|P_i, D_i = 1] = E[E[W_i|X_i]|P_i, D_i = 1] \\ &= \frac{E[D_i E[W_i|X_i]|P_i]}{E[D_i|P_i]} = \frac{E[P_i E[W_i|X_i]|P_i]}{P_i} \\ &= E[E[W_i|X_i]|P_i] = E[W_i|P_i]. \end{aligned}$$

By similar reasoning we also have $E[W_i|P_i, D_i = 0] = E[W_i|P_i]$, so the conclusion follows by the previous equation.

## Regression Discontinuity Design

There are two cases here, one where the treatment variable jumps discontinuously, and one where the treatment probability is discontinuous. Consider the discontinuous treatment variable first.

We suppose that $D_i = 1(X_i \geq c)$. In this case $E[Y_{i0}|D_i, X_i] = E[Y_{i0}|X_i]$ and $E[Y_{i1}|D_i, X_i] = E[Y_{i1}|X_i]$ hold by construction. The common support assumption is not satisfied. Indeed, in this case $\mathcal{X}_0$ and $\mathcal{X}_1$ are disjoint.

We take a different approach to identification here, and instead rely on a continuity condition for $E[Y_{i0}|X_i = x]$ and $E[Y_{i1}|X_i = x]$.

Assumption: $E[Y_{i0}|X_i = x]$ and $E[Y_{i1}|X_i = x]$ are continuous in $x$ at $c$.

Note that for $Y_i = Y_{i0}$ for $X_i < c$ and $Y_i = Y_{i1}$ for $X_i \geq c$. Then

$$\begin{aligned} E[Y_{i0}|X_i &= c] = \lim_{x \uparrow c} E[Y_{i0}|X_i = x] = \lim_{x \uparrow c} E[Y_i|X_i = x], \\ E[Y_{i1}|X_i &= c] = \lim_{x \downarrow c} E[Y_{i1}|X_i = x] = \lim_{x \downarrow c} E[Y_i|X_i = x]. \end{aligned}$$

It follows that

$$E[\beta_i|X_i = c] = E[Y_{i1} - Y_{i0}|X_i = c] = \lim_{x \downarrow c} E[Y_i|X_i = c] - \lim_{x \uparrow c} E[Y_i|X_i = c].$$

Thus, the conditional treatment effect $E[Y_{i1} - Y_{i0}|X_i = c]$ is identified as the jump in $E[Y_i|X_i = x]$ at $x = c$.

We can also interpret this differently. Similarly to above,

$$E[\beta_i|X_i = c] = E[Y_i|D_i = 1, X_i = c] - E[Y_i|D_i = 0, X_i = c].$$

Note that both of $E[Y_i|D_i = 1, X_i = c]$ and $E[Y_i|D_i = 0, X_i = c]$ are nonparametric regression functions evaluated at the boundary of their support, the first at the lower boundary and the second at the upper. So regular kernel regression is not good. Can do locally linear regression.