

# Average Treatment Effect

Li Gan

September 2009

We are interested in average changes in outcome  $y$  after a policy change. As in the case of the medical field, we are interested in the outcomes of a new medicine, including its effectiveness and its side effect. To do that, it is typical that we randomly divide patients into two groups, the treatment group, and the control group. The treatment group is given the medicine while the control group is given the “placebo”.

Consider an economic example. An important policy question is how to help needy families. Income transferring programs, such as Aid to Families with Children (AFDC) creates disincentives of working. One alternative method is the Earned Income Tax Credit (EITC).

Eligibility for EITC: Gross income below a specified amount (in 2007, the amount is \$39,783 if you children and \$14,590 if you do not have children).

Benefits (in 2007): maximum benefits: \$428 if no children; \$2,853 if one child; and \$4,716 if two children.

The Tax Reform Act of 1986 includes an expansion of earned income tax credit. We are interested if expansion of EITC helps increasing labor supply.

Denote 1 if with treatment (experiences expansions of EITC), and 0 without treatment (not affected by expansions of EITC). Average Treatment Effect ( $ATE$ ) is defined as:

$$ATE = E(y_1 - y_0) \tag{1}$$

The difficulty in estimating (1) is that we observe either  $y_1$  or  $y_0$ , not both, for each person.

## 1. Regression method: the *Difference-in-difference method*

In this case, we potentially can observe outcomes before the treatment and after the treatment for the same person or for different persons. We have two time periods, say year 0 and year 1. One would say that we can simply apply (1). In the case of EITC expansion, year 0 is before 1986, and year 1 is after 1986.

However, this is not entirely appropriate since there may be other factors that affect treated people as well. The difference in labor supply before 1986 and after 1986 may be due to overall economic environment.

Therefore, we need a control group. There are two groups, the control group (denoted as group A), and the treatment group (denoted as B). At period 0, no treatment for both groups. At period 1, the treatment group experiences policy change (treatment) while the control group does not. Let  $D_1$  denote a dummy variable for time period 1, and  $D_B$  denote the treatment group. The simplest regression for analyzing the impact of the policy change is:

$$y = \beta_0 + \delta_0 D_1 + \beta_1 D_B + \delta_1 D_1 * D_B + u \quad (2)$$

It is easy to show that:

$$\hat{\delta}_1 = (\bar{y}_{B,1} - \bar{y}_{B,0}) - (\bar{y}_{A,1} - \bar{y}_{A,0}) \quad (3)$$

This is why the regression of (2) is often called the difference-in-difference.

In the example of the expansions of EITC, Eissa and Liebman use “*single women without children*” as the control group, and “*single women with children*” as the treatment group. Their time periods are: 1984-1986 as time 0, and 1988-1990 as time 1.

The regression, therefore, is:

$$\Pr(lfp_{it} = 1) = \Phi(\alpha + \beta Z_{it} + \gamma_0 ChildrenDummy_i + \gamma_1 post86_t + \gamma_2 (ChildrenDummy_i \times post86_t))$$

Eissa and Liebman (1996) find that single women with children increased their relative labor force participation by up to 2.8% percentage points.

## 2. Regression method

More generally, suppose we do not observe outcome variables  $y$  before and after treatment. Then: let  $w = 1$  if treatment. The observed outcome  $y$  can be written as:

$$y = (1-w) y_0 + w y_1. \quad (4)$$

If  $w$  is independent of  $y$ , then:

$$E(y_1 - y_0) = E(y_1 - y_0 | w) = E(y_1 | w=1) - E(y_0 | w=0)$$

In fact, we only need the weaker assumption (rather than independence): mean independence:  $E(y_0 | w) = E(y_0)$ ,  $E(y_1 | w) = E(y_1)$ .

In the Eissa and Liebman (1996) example, the treatment group is the “*single women with children*” while the control group is the “*single women without children*.” It is entirely possible the treatment variable is NOT independent with the outcome variable (labor force participation).

Now let:

$$\begin{aligned} y_0 &= \mu_0 + v_0, & E(v_0) &= 0 \\ y_1 &= \mu_1 + v_1, & E(v_1) &= 0 \end{aligned}$$

Therefore, (4) can be written as:

$$\begin{aligned} y &= (1-w) y_0 + w y_1 \\ &= \mu_0 + (\mu_1 - \mu_0) w + v_0 + w (v_1 - v_0) \end{aligned} \quad (5)$$

First, assume conditional mean independence:

*Assumption 1* (ATE 1): (a)  $E(y_0|w,x) = E(y_0|x)$ , and (b)  $E(y_1|w,x) = E(y_1|x)$

Intuition: even though  $y_1$  and  $y_0$  may be correlated with  $w$ , they are uncorrelated with  $w$  if we partial out  $x$ .

Taking expectation of (5) (and with *ATE 1*):

$$E(y|w,x) = \mu_0 + \alpha w + g_0(x) + w(g_1(x) - g_0(x)), \quad (6)$$

where  $\alpha = \mu_1 - \mu_0$  is the *ATE*, and  $g_i(x) = E(v_i|x)$ .

Linearization of  $g_i(x)$ :  $g_i(x) = x\beta_i$ .

$$E(y|w,x) = \mu_0 + \alpha w + x\beta_0 + wx(\beta_1 - \beta_0).$$

Rewrite it:

$$E(y|w,x) = \mu_0 + \alpha w + x\beta_0 + w(x - \psi)\delta,$$

where  $\psi = E(x)$ , and  $\delta = \beta_1 - \beta_0$ . The last term is to ensure that  $g_1(x) - g_0(x) = 0$ . So the regression to estimate *ATE*  $\alpha$  is:

$$y_i \text{ on } 1, w_i, x_i, w_i(x_i - \bar{x}) \quad (6-1)$$

Here the control functions involve not just  $x_i$ , but also the interactions of the covariates with the treatment variable.

We can estimate treatment effect conditional on  $x$ :

$$\hat{ATE}(x) = \hat{\alpha} + (x - \bar{x})\hat{\delta}$$

Discussions:

Compare the model in (6-1) and the difference-in-difference estimator of (2), the difference is that (6-1) allows the coefficients for  $x$  to be different between the treated group ( $w_i = 1$ ) and the non-treated group ( $w_i = 0$ ). This can be seen more clearly that (6-1) assumes that:

$$E(v_1|x_i) \equiv g_1(x_i) \neq g_0(x_i) \equiv E(v_0|x_i).$$

### 3. Propensity Score:

Let  $p(x) = \Pr(w=1|x)$ .

$$\begin{aligned} (w - p(x))y &= (w - p(x))(wy_1 + (1-w)y_0) \\ &= wy_1 - p(x)(1-w)y_0 - p(x)wy_1 \end{aligned}$$

Take conditional expectation with respect to  $y$ :

$$E_y[(w - p(x))y|w,x] = wm_1(x) - p(x)(1-w)m_0(x) - p(x)wm_1(x),$$

where  $E(y_j|w,x) = E(y_j|x) = m_j(x)$ . Taking expectation with respect to  $w$ :

$$\begin{aligned} &E_w\{E_y[(w - p(x))y|w,x]|x\} \\ &= E_w[wm_1(x) - p(x)(1-w)m_0(x) - p(x)wm_1(x)] \\ &= p(x)m_1(x) - p(x)(1-p(x))m_0(x) - p(x)p(x)m_1(x) \\ &= m_1(x)p(x)(1-p(x)) - m_0(x)p(x)(1-p(x)) \\ &= (m_1(x) - m_0(x))p(x)(1-p(x)) \end{aligned}$$

Therefore,

$$\begin{aligned} ATE &= m_1(x) - m_0(x) \\ &= \frac{E((w - p(x))y)}{p(x)(1 - p(x))} \end{aligned}$$

A simple and popular estimator in program evaluation is obtained from OLS regression:

$$y_i \text{ on } 1, w_i, \hat{p}(x_i)$$

where coefficient for  $w_i$  is the estimate of the treatment effect. In other words, the estimated propensity score plays the role of the control function.

### 4. Dummy Endogenous Variables

Consider the model again:

$$E(y|w, x) = \mu_0 + \alpha w + x\beta_0 + u_0, \quad (7)$$

$w$  is endogenous. Again,  $w = 1$  if treated, and 0 otherwise.

Assume that  $\Pr(w=1|x,z) = G(x, z; \gamma)$

Procedure 1:

- (1) Estimate the binary response model  $\Pr(w_i=1|x_i,z_i) = G(x_i,z_i;\gamma)$ , and obtain the fitted values  $\hat{G}_i$ .
- (2) Estimate (4) using instruments 1,  $\hat{G}_i$  and  $x_i$ .

Procedure 1 has important robustness property:

- (a) Because we use  $\hat{G}_i$  as an *IV*, the model  $\Pr(w_i=1|x_i,z_i) = G(x_i,z_i;\gamma)$  does not have to be correctly specified.
- (b) Technically,  $\alpha$  and  $\beta$  are identified even if we do not have extra variables excluded from  $x$ . In other words, we do not have  $z_i$ . In this case, the identification is completely coming from nonlinear function of  $x$ .

However, we can rarely justify the estimator in this case. Suppose that  $w$  given  $x$  follows a probit model (no  $z$ ). Because  $G(x, \gamma) = \Phi(\gamma_0 + x\gamma_1)$ , is a nonlinear function of  $x$ , it is not perfectly correlated with  $x$ , so it can be used as IV for  $w$  technically.

- (c) In principle, it important to recognize that Procedure 1 is not the same as using  $G$  as a regressor in place of  $w$ .

$$y_i \text{ on } 1, \hat{G}_i \text{ and } x_i.$$

Consistency of the OLS estimators from the regression:

$$y_i = \delta_0 + \alpha\hat{G}_i + x_i\beta_0 + u_i \quad (8)$$

relies on  $G(\cdot)$  to be correctly specified. Note that (8) also has problems with standard errors that need to be corrected.

Further, if we allow the interaction term:

$$y_i = \delta_0 + \alpha w_i + x_i\beta_0 + w_i(x_i - \bar{x})\delta + e_i \quad (9)$$

Procedure 2:

- (a) Estimate  $\Pr(w_i=1|x_i,z_i) = G(x_i,z_i;\gamma)$
- (b) Use 1,  $\hat{G}_i$  and  $x_i$ , and  $\hat{G}_i(x_i - \bar{x})$  as IVs.

Discussions are the same as before.

## 5. Regression discontinuity

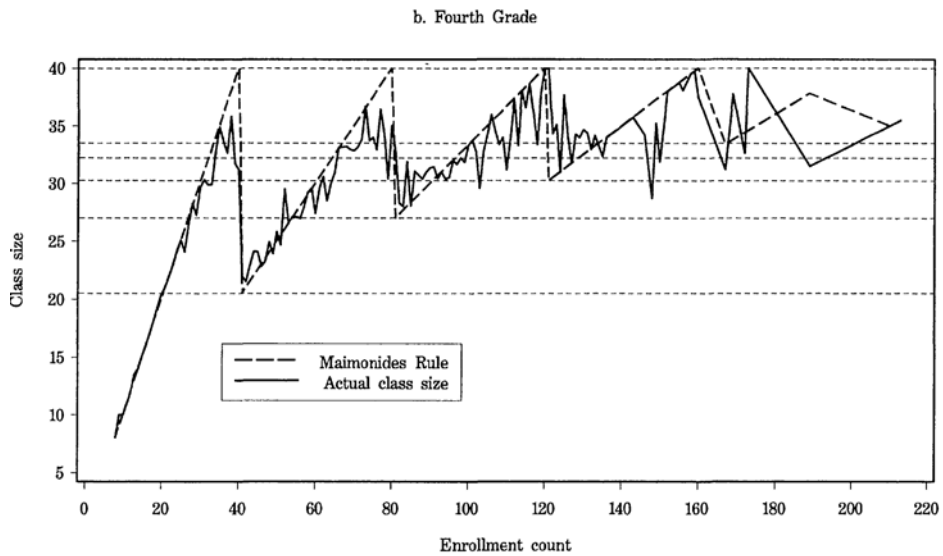
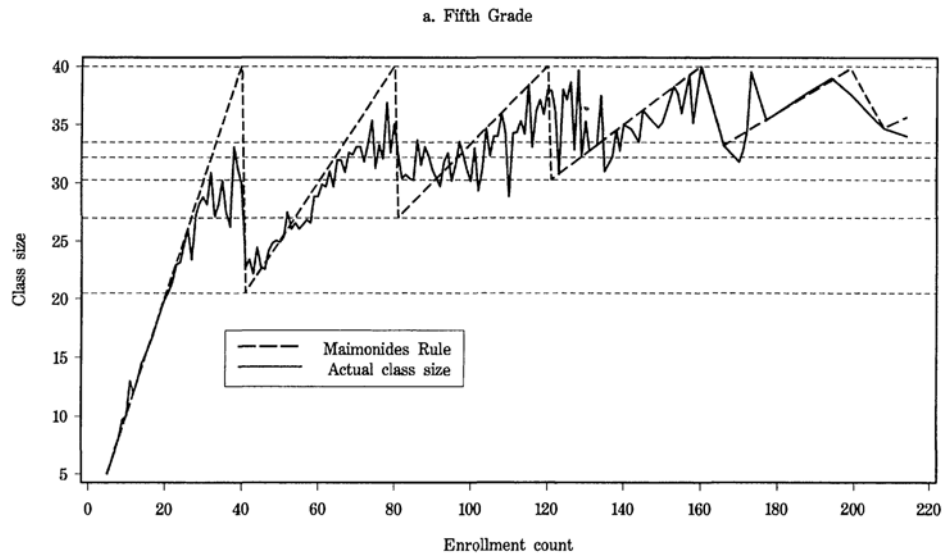
Consider an example: we are interested in effect of class size on the performance of students. A typical assumption is that a smaller class size helps learning.

The problem of a typical study, obviously, is that the class size is not exogenous. It is often determined partly by the student performance.

In Angrist and Lavy (Quarterly Journal of Economics, 1999. *Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement*), suppose there is a law requiring class size to be less than 40<sup>1</sup>; if a school has 120 students, then there will be four classes. However, if a school has 121 students, then there would be five classes. The class size is smaller for schools with 121 students than schools for 120 students. These two schools should be similar in every aspect other than class size.

---

<sup>1</sup> This is due to the great twelfth century Rabbinic scholar, Maimonides. According to Angrist and Lavy (1999), the Maimonides says, "Twenty-five children may be put in charge of one teacher. If the number in the class exceeds twenty-five but is not more than forty, he should have an assistant to help with the instruction. If there are more than forty, two teachers must be appointed."



Class Size in 1991 by Initial Enrollment Count, Actual Average Size and as Predicted by Maimonides' Rule

Other examples:

- Anti-poverty program → targeted to households below a given poverty index.
- Pension program → targeted to population above a certain age.
- Tax rebate → income lower than a certain threshold.

It is useful to distinguish between two general settings, the Sharp and the Fuzzy Regression Discontinuity designs.

*Sharp Design:*

$$w_i = 1(x_i > x_0)$$

The assignment  $w_i$  is a deterministic function of one of the covariates, the forcing (or treatment-determining) variable  $x$ .

All units with  $x_i > x_0$  are assigned to the treatment group (and participation is mandatory for these individuals), and all units with  $x_i \leq x_0$  are assigned to the control group. In this sharp design, we look at the discontinuity in the conditional expectation of the outcome given the covariates to uncover the *ATE*:

$$ATE = \lim_{x \rightarrow x_0^+} E[y | x] - \lim_{x \rightarrow x_0^-} E[y | x] = E(y_1 - y_0 | x = x_0)$$

*Fuzzy design:*

$E(w_i | x_i = x) = \Pr(w_i = 1 | x)$  is discontinuous at known value  $x_0$ . In this case, it is observed that if a person is treated or not. ( $w_i$  is observed). However, it is no longer the case that all  $x_i > x_0$  belongs to the treatment group. Some  $x_i < x_0$  will have  $w_i = 1$ .

Assumption: (i)  $w^+ = \lim_{x \rightarrow x_0^+} E(w | x)$  and  $w^- = \lim_{x \rightarrow x_0^-} E(w | x)$  exist.  
(ii)  $w^+ \neq w^-$

The sharp and fuzzy designs differ in that in the sharp design the treatment assignment is deterministic given  $x$ , while the fuzzy design the treatment assignment may depend on additional factors unobserved by econometrician. In both designs, the discontinuity point  $x_0$  is known.

In Angrist and Lavy (1999), it is a sharp design, with known  $x_0$  being at the multiples of 40, i.e.,  $x_0 = 40, 80, 120, \dots$

Assumption:  $E(y_{1i} - y_{0i} | x_i = x)$  is continuous in  $x$  at  $x_0$ .

This assumption is valid where we have reason to believe that person close to threshold  $c$  are similar and thus would experience similar outcome absent treatment.

Theorem: *ATE*, denoted as  $\alpha$ :

$$\alpha = \frac{y^+ - y^-}{w^+ - w^-}$$

Proof:

Let  $\Delta$  to be a small positive number.

$$\begin{aligned}
& E(y | x_0 + \Delta) - E(y | x_0 - \Delta) \\
&= E((y_1 - y_0)w + y_0 | x_0 + \Delta) - E((y_1 - y_0)w + y_0 | x_0 - \Delta) \\
&= E((y_1 - y_0)w | x_0 + \Delta) - E((y_1 - y_0)w | x_0 - \Delta) + (E(y_0 | x_0 + \Delta) - E(y_0 | x_0 - \Delta)) \\
&= \alpha(E(w | x_0 + \Delta) - E(w | x_0 - \Delta)) + (E(y_0 | x_0 + \Delta) - E(y_0 | x_0 - \Delta))
\end{aligned}$$

As  $\Delta \rightarrow 0$ , we have:

$$y^+ - y^- = \alpha(w^+ - w^-)$$

Here we use the fact (assumption) that  $E(y_0)$  is continuous at  $x_0$  without treatment. The conclusion follows.

Given this theorem, we can obtain an estimate of  $\alpha$  by estimating  $y^+$ ,  $y^-$ ,  $w^+$ , and  $w^-$ .

There are several ways to estimate this. The most popular way is to do it non-parametrically.

In practice,

$$\begin{aligned}
\hat{y}^+ &= \frac{\sum y_i 1(x_0 < x_i < x_0 + h)}{\sum 1(x_0 < x_i < x_0 + h)} \\
\hat{y}^- &= \frac{\sum y_i 1(x_0 - h < x_i < x_0)}{\sum 1(x_0 - h < x_i < x_0)}
\end{aligned}$$

Note for a sharp design RD,  $w^+ - w^- = 1$ . For a fuzzy design RD,

$$\begin{aligned}
\hat{w}^+ &= \frac{\sum w_i 1(x_0 < x_i < x_0 + h)}{\sum 1(x_0 < x_i < x_0 + h)} \\
\hat{w}^- &= \frac{\sum w_i 1(x_0 - h < x_i < x_0)}{\sum 1(x_0 - h < x_i < x_0)}
\end{aligned}$$

where  $h$  is the bandwidth. An interesting note is that this is numerically equivalent to an IV estimator for the regression:

$y_i$  on  $w_i$  for people in the subsample  $(x_0 - h < x_i < x_0 + h)$   
using  $1(x_0 < x_i < x_0 + h)$  as the IV.

The regression method can be useful because one can add control variables in the regression. Note for the fuzzy design, it is not necessary that  $1(x_0 < x_i < x_0 + h) = 1$ .

Practically, for a sharp design,

1. Graph the data by computing the average value of the outcome variable over a set of bins. The bandwidth has to be large enough to have sufficient amount of precision so that the plots look smooth on either side of the cutoff value, but at the same time small enough to make the jump around the cutoff value clear.
2. Estimate the treatment effect by running linear regression on both sides of the cutoff point. Since we propose to use a rectangular kernel, these are just standard regression estimated within a bin of width  $h$  on both sides of cutoff point. Note that:
  - i. Standard errors can be computed using standard least square methods (robust standard errors)
  - ii. The optimal bandwidth can be chosen using cross-validation methods.

Fuzzy design:

1. Graph the average outcome over a set of bins as in the case of SRD, but also graph the probability of treatment.
2. Estimating the treatment effect using TSLS.

Figures 1 and 2 shows that without treatment,  $E(y_0 | x = x_0^+) = E(y_0 | x = x_0^-)$ . Here  $x_0$  is 50.

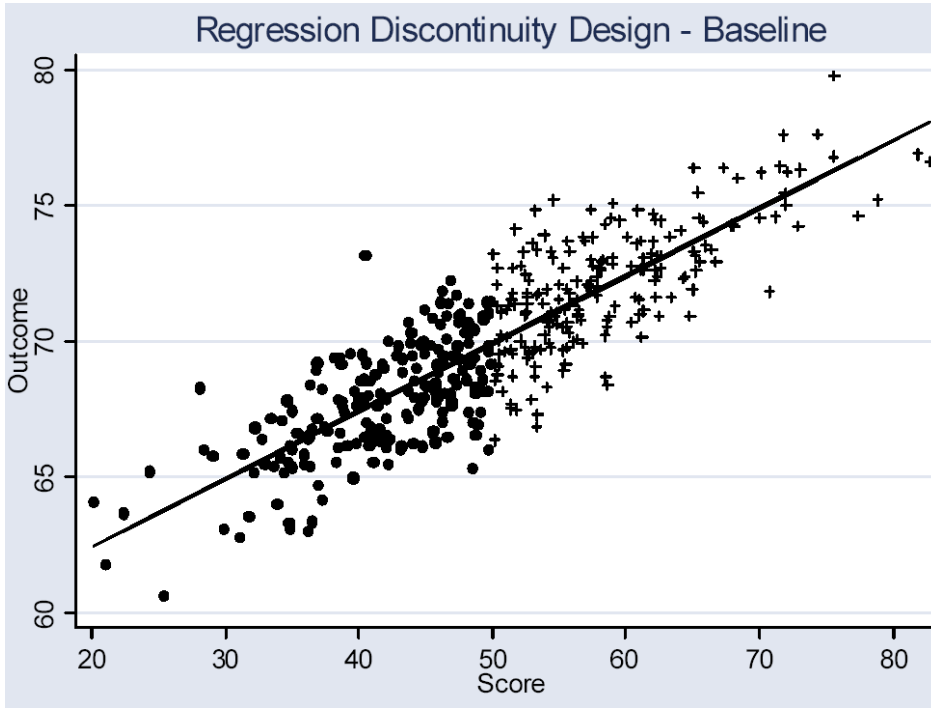


Figure 2

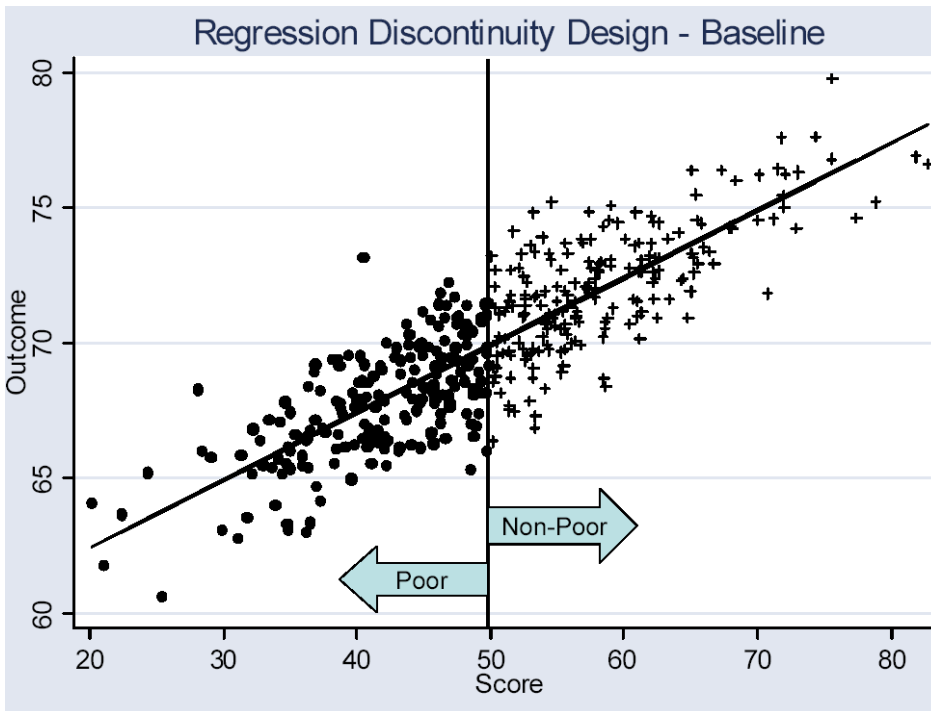


Figure 3 shows that the effect of the treatment.  $E(y_0 | x = x_0^+) < E(y_0 | x = x_0^-)$

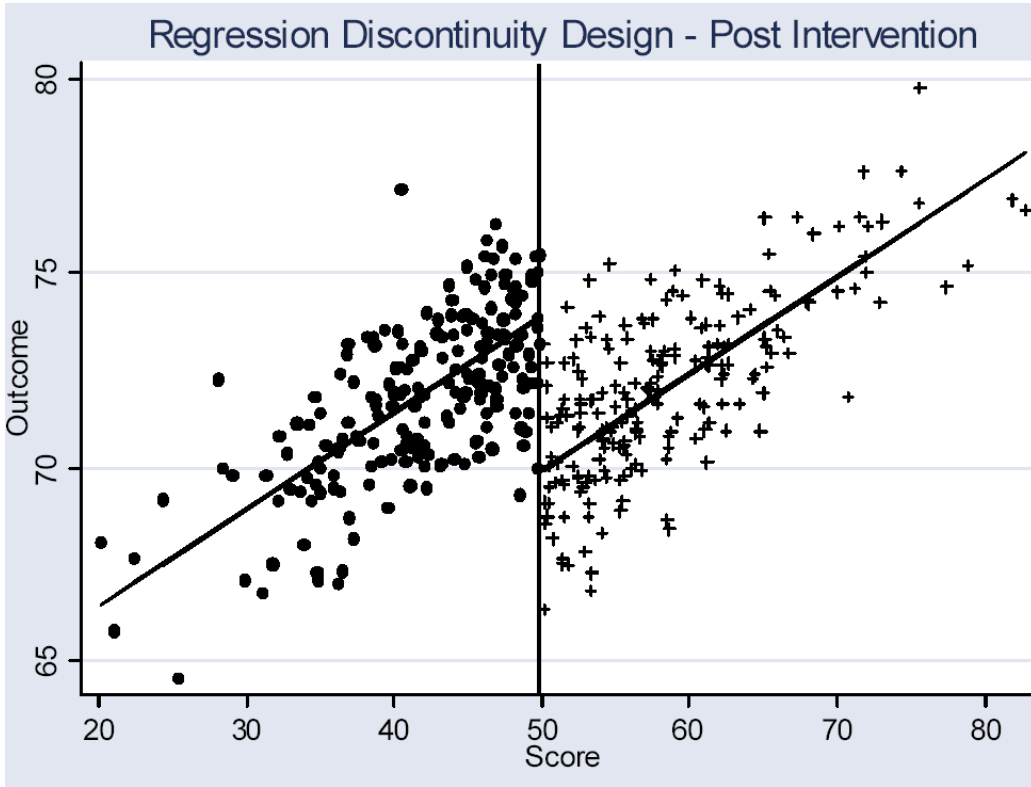


Figure 4 shows the Average Treatment Effect.

