Chapter 2. Social Experiments (Randomized Control Trials and Natural Experiments)

Joan Llull

Quantitative & Statistical Methods II — Part I
Barcelona School of Economics

I. Randomized Control Trials and Natural Experiments

In the treatment effect approach, a ***randomized field trial*** is regarded as the ideal research design. Observational studies are seen as more speculative attempts to generate the force of evidence of experiments.

There is a long history of randomized field trials in social welfare in the U.S., beginning in the 1960s (see Moffitt, 2003, for a review). Early experiments had many flaws due to the lack of experience in designing them, and in data analysis. During the 1980s, the U.S. Federal Government started to encourage states to use experimentation, eventually becoming almost mandatory. The analysis of the 1980s experimental data consisted of simple treatment-control differences. The force of the results had a major influence on the 1988 legislation. In spite of these developments, randomization encountered resistance from many U.S. states on ethical grounds. Even more so in other countries, where treatment groups have often been formed by selecting areas for treatment instead of individuals.

Experiments are often very expensive, and often difficult to implement. However, nature sometimes do the job, providing ***natural experiments***. Very illustrative to this end is the way in which science connected cholera and the quality of drinking water in the SoHo in London, in 1854. In the 19th century, London suffered from periodic cholera epidemics in which many died. Cholera was believed to be caused by bad air quality, but John Snow (a medical doctor) suspected that instead it was caused by bad water quality (though he had no theory of why). In order to use experimental data to test this hypothesis, one could randomly give some people good water and some people bad water. However, there are good ethical reasons why this experiment cannot be implemented on people.

In 1854 there was a severe outbreak of cholera in Soho. Snow thought contamination of the pump in Broad Street was the source of the problem. He found those for whom this was the closest pump were more likely to die, but in nearby workhouse fewer people died (they had their own well). The brewery on Broad Street itself reported no deaths (they also had their own well —though the men normally only drank beer). These two groups breathed the same air but had ac-

cess to different water. A further piece of evidence was two isolated deaths, one in Hampstead, one in Islington, of an aunt and her niece. The aunt was in the habit of having a barrel of water delivered from the Broad Street pump every day (she liked the taste) and the niece had paid her a visit. Thus, even though this variation in who drank what water was not assigned at random by any researcher, he built a powerful case that "bad water" was the source of problem. This accidental variation in the source of water can be considered "as good as randomly assigned", because of different water sources across houses in the same street and sometimes even across apartments within houses. Some houses got their drinking water supply from companies such as Lambeth that sourced upstream, i.e. above sewage discharge points, while other houses were supplied by companies such as Southwark and Vauxhall that sourced downstream, i.e. from dirtier water. Snow identified the water companies for the houses with cholera deaths as well as the total number of houses served by each company in his study area. And results corroborated his suspicion.

II. Random Assignment and Treatment Effects

In a controlled experiment, treatment status is randomly assigned by the researcher, which by construction, ensures independence:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i. \tag{1}$$

As noted in Chapter 1, this eliminates the selection bias (and implies $\alpha_{TT} = \beta$), as:

$$\mathbb{E}[Y_{0i}|D_i = 1] = \mathbb{E}[Y_{0i}|D_i = 0] = \mathbb{E}[Y_{0i}]. \tag{2}$$

It also implies $\alpha_{ATE} = \alpha_{TT} = \beta$, as $\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i}]$. Thus, the average treatment effect can be estimated by a simple linear regression of the observed outcome Y_i on the treatment dummy D_i and a constant.

III. Standard Errors and Inference

When implementing this estimation as a linear regression with a standard statistical package (say, Stata or R) the default options provide an estimate of the standard error of the slope coefficient β, which gives, in this context, an estimate of the standard error of the estimated average treatment effect. The standard default options usually assume that residuals U_i are homoskedastic. This implies:

$$\text{Var}(Y_i|D_i = 1) = \text{Var}(Y_i|D_i = 0) = \text{Var}(U_i) = \text{Var}(Y_{0i}). \tag{3}$$

This assumption is often violated in the context of heterogeneous treatment effects. Noting that $\text{Var}(Y_i|D_i = 1) = \text{Var}(Y_{1i}|D_i = 1)$ by definition, and that, in this context, given independence, $\text{Var}(Y_{1i}|D_i = 1) = \text{Var}(Y_{1i})$, Equation (3) implies $\text{Var}(Y_{1i}) = \text{Var}(Y_{0i})$. However, $\text{Var}(Y_{1i})$ can be expressed as:

$$\text{Var}(Y_{1i}) = \text{Var}(Y_{0i}) + \text{Var}(Y_{1i} - Y_{0i}) + 2\,\text{Cov}(Y_{1i} - Y_{0i}, Y_{0i}), \tag{4}$$

which is obtained by noting that $\text{Var}(Y_{1i}) = \text{Var}(Y_{0i} + (Y_{1i} - Y_{0i}))$. In the context of homogeneous treatment effects, $Y_{1i} - Y_{0i}$ is a constant, and, thus, the second and third terms of the right hand side are equal to zero. But, in general, this is not the case with heterogeneous treatment effects.

For example, consider the case in which treatment effects are heterogeneous (and hence $\text{Var}(Y_{1i} - Y_{0i}) > 0$), but uncorrelated (or positively correlated) with the initial outcome. For example, consider the effect of buying a lottery ticket on wealth. To analyze that, the researcher randomly provides lottery tickets to subjects independently of their wealth. Among treated individuals, with probability p their wealth is increased by an amount M, and with probability $1-p$, the wealth increases by 0. In this context:

$$\text{Var}(Y_{1i} - Y_{0i}) = M^2 \cdot p + 0 \cdot (1 - p) = pM^2 > 0, \tag{5}$$

and $\text{Cov}(Y_{1i} - Y_{0i}, Y_{0i}) = 0$ as both the prize M and the probability of winning p are independent of the initial wealth Y_{0i}. In this case, $\text{Var}(Y_{1i}) > \text{Var}(Y_{0i})$, and the homoskedasticity assumption is violated.

This violation does not affect the bias and consistency of the OLS estimator. But it implies that the default standard errors may be inconsistent. Thus, it is useful to provide alternatives to the estimation of standard errors. If observations in the sample are independent from each other, a natural way to compute the standard error of the average treatment effect is by focusing on the difference in means rather than in the regression estimation, and noting that:

$$\text{Var}(\beta^S) = \text{Var}(\bar{Y}_T - \bar{Y}_C) = \text{Var}(\bar{Y}_T) + \text{Var}(\bar{Y}_C) = \frac{\sigma_T^2}{N_1} + \frac{\sigma_C^2}{N_0}, \tag{6}$$

where σ_T^2 and σ_C^2 are respectively the variances of the outcome computed on treated and control subsamples, and N_1 and N_0 are the sizes of each subsample, as defined in Chapter 1. The second equality makes use of the independence across observations. A sample analog could be computed with sample variances or corrected sample variances.

An alternative way to obtain standard errors when observations are independent follows the regression approach and computes ***robust*** standard errors. Recall

from QSM I that the asymptotic formula for the variance of the estimators under heteroskedasticity is given by the following sandwich formula:

$$\text{Var}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}\end{pmatrix} = \frac{1}{N}\,\mathbb{E}[X_iX_i']^{-1}\,\mathbb{E}[X_iX_i'U_i^2]\,\mathbb{E}[X_iX_i']^{-1}, \tag{7}$$

where $X_i \equiv (1, D_i)'$. The sample analog of the above expression is:

$$\widehat{\text{Var}}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}\end{pmatrix} = \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i'\hat{U}_i^2\right)\left(\frac{1}{N}\sum_{i=1}^{N}X_iX_i'\right)^{-1}. \tag{8}$$

This expression is general for any vector of regressors X_i, but in this context it simplifies further. To check that, we need to operate the different matrices and sums. The derivation below uses the fact that $D_i^2 = D_i$, as discussed in Chapter 1:

$$\begin{aligned}
\widehat{\text{Var}}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}\end{pmatrix} &= \frac{1}{N}\left[\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}1 & D_i\\ D_i & D_i\end{pmatrix}\right]^{-1}\left[\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}1 & D_i\\ D_i & D_i\end{pmatrix}\hat{U}_i^2\right]\left[\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}1 & D_i\\ D_i & D_i\end{pmatrix}\right]^{-1}\\
&= \frac{1}{N}\left[\begin{pmatrix}1 & \frac{N_1}{N}\\ \frac{N_1}{N} & \frac{N_1}{N}\end{pmatrix}\right]^{-1}\begin{pmatrix}\frac{\sum_{i=1}^{N}\hat{U}_i^2}{N} & \frac{\sum_{i:D_i=1}\hat{U}_i^2}{N}\\ \frac{\sum_{i:D_i=1}\hat{U}_i^2}{N} & \frac{\sum_{i:D_i=1}\hat{U}_i^2}{N}\end{pmatrix}\left[\begin{pmatrix}1 & \frac{N_1}{N}\\ \frac{N_1}{N} & \frac{N_1}{N}\end{pmatrix}\right]^{-1}\\
&= \left[\begin{pmatrix}N & N_1\\ N_1 & N_1\end{pmatrix}\right]^{-1}\begin{pmatrix}\sum_{i=1}^{N}\hat{U}_i^2 & \sum_{i:D_i=1}\hat{U}_i^2\\ \sum_{i:D_i=1}\hat{U}_i^2 & \sum_{i:D_i=1}\hat{U}_i^2\end{pmatrix}\left[\begin{pmatrix}N & N_1\\ N_1 & N_1\end{pmatrix}\right]^{-1}\\
&= \begin{pmatrix}\frac{1}{N_0} & -\frac{1}{N_0}\\ -\frac{1}{N_0} & \frac{1}{N_0}+\frac{1}{N_1}\end{pmatrix}\begin{pmatrix}\sum_{i=1}^{N}\hat{U}_i^2 & \sum_{i:D_i=1}\hat{U}_i^2\\ \sum_{i:D_i=1}\hat{U}_i^2 & \sum_{i:D_i=1}\hat{U}_i^2\end{pmatrix}\begin{pmatrix}\frac{1}{N_0} & -\frac{1}{N_0}\\ -\frac{1}{N_0} & \frac{1}{N_0}+\frac{1}{N_1}\end{pmatrix}\\
&= \begin{pmatrix}\frac{\sum_{i:D_i=0}\hat{U}_i^2}{N_0} & 0\\ \frac{\sum_{i:D_i=1}\hat{U}_i^2}{N_1}-\frac{\sum_{i:D_i=0}\hat{U}_i^2}{N_0} & \frac{\sum_{i:D_i=1}\hat{U}_i^2}{N_1}\end{pmatrix}\begin{pmatrix}\frac{1}{N_0} & -\frac{1}{N_0}\\ -\frac{1}{N_0} & \frac{1}{N_0}+\frac{1}{N_1}\end{pmatrix}\\
&= \begin{pmatrix}\frac{\sum_{i:D_i=0}\hat{U}_i^2}{N_0^2} & -\frac{\sum_{i:D_i=0}\hat{U}_i^2}{N_0^2}\\ -\frac{\sum_{i:D_i=0}\hat{U}_i^2}{N_0^2} & \frac{\sum_{i:D_i=0}\hat{U}_i^2}{N_0^2}+\frac{\sum_{i:D_i=1}\hat{U}_i^2}{N_1^2}\end{pmatrix},
\end{aligned} \tag{9}$$

where the fourth row uses the identity $N = N_0 + N_1$ and fifth and sixth use $\sum_{i=1}^{N}\hat{U}_i^2 = \sum_{i:D_i=0}\hat{U}_i^2 + \sum_{i:D_i=1}\hat{U}_i^2$. Thus, the estimated variance of the average treatment effect (bottom right quantity) is of the form of sum of sample variances of treated and control errors divided by the corresponding sample sizes, as $\hat{U}_i = Y_i - \bar{Y}_C$ for control observations and $U_i = Y_i - \bar{Y}_T$ for treated ones. In other words, the regression robust standard error provides an estimate that is numerically equivalent to the variance of the average treatment effect computed directly.

In all this derivation, we assumed that observations are independent. However, the experimental design and data collection sometimes generates correlation between a subset of observations. For example, in the Progresa conditional cash

transfers introduced in Mexico to foster children's education, randomization was done at the village level, and, within a village, all individuals were treated or controls. This implies that some observations experience common village-level shocks, and, thus, are correlated. A solution to this is ***clustering*** standard errors (at the village level in this case). We will expand on this aspect further if time permits at the end of the course.

IV. Introduction of Additional Regressors

The discussion above shows that econometrics would be very easy if all data was from (well executed) randomized control experiments: one could get causal effects simply by comparing means, there would be no need for matrix algebra or even multiple regressions, and no need to collect any variables other than treatment status and the outcome variable.

However, even in this setup, there are situations in which additional regression can be useful. Let W_i denote a vector of additional possible regressors. Randomization ensures consistency, even if they are not included. The omitted variable bias formula obtained in Equation (29) in Chapter 1 is:

$$\gamma \frac{\text{Cov}(W_i, D_i)}{\text{Var}(D_i)}, \tag{10}$$

which is equal to zero because randomization implies $\text{Cov}(W_i, D_i) = 0$. This is so unless W_i is a "bad control", that is, an intermediate outcome that is affected by the treatment (as occupational choice in Chapter 1). We revisit this point below.

One advantage of including additional controls is that, if they are relevant, this would typically increase precision in the estimated average treatment effect. Intuitively this is so because by holding constant other characteristics that affect the outcomes, we are reducing the variance of U_i. More formally, one can apply the partial regression results by Frisch and Waugh to show it. The Frisch-Waugh Theorem (whose proof is quite straightforward but out of the scope of the course, so you can easily check in any textbook or even online) establishes that if we are interested in β_1 in the following regression:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + U_i, \tag{11}$$

we can apply two different procedures that provide exactly (numerically) the same result. The first one is OLS on the whole regression. The second is to regress X_{1i} and Y_i on X_{2i}, obtain the residuals of the two regressions, namely V_i and E_i

respectively, and then estimate the following regression:

$$E_i = \beta_1 V_i + U_i. \tag{12}$$

Using this result, and noting that, since the regression of W_i on the treatment variable D_i provides a zero coefficient given independence (through randomization), the resulting regression would be:

$$E_i = Y_i - \gamma_Y W_i = \tilde{\beta}_0 + \beta D_i + \tilde{U}_i, \tag{13}$$

where γ_Y is the regression coefficient of Y_i on W_i, and $\tilde{\beta}_0 \equiv \mathbb{E}[Y_{0i} - \gamma_Y W_i]$. Since $\text{Var}(E_i) \leq \text{Var}(Y_i)$, then $\text{Var}(\tilde{U}_i) \leq \text{Var}(U_i)$.

A different role for additional regressors in randomized experiments is related to checking for randomization. In many real world examples of randomized experiments, there are serious questions about how well the randomization was implemented in practice. In such situation, adding some controls can provide additional credibility to the results. A previous step to that in practice is to check whether the possible additional controls have different averages in treated and control groups. A fast way of doing it is to compute the treatment effect on these variables and test whether we can reject that it equals zero. If it appears that treatment and control samples differ in a particular dimension, including this variable as a control could eliminate the resulting omitted variable bias.

Similarly, sometimes randomization is implemented conditional on observables. Regressors can be useful at the design stage. For example, in the Progresa case, randomization is at the village level. This ensures that control units are not "contaminated" by treatment of treated units. In these cases, we need to further control for the variables used in the randomization design. This and the previous cases lead to the conditional independence situation, discussed in the next chapter.

V. Warnings: Partial or Imperfect Compliance and Longer Run Interaction of Treatment and Intermediate Outcomes

A. Partial or Imperfect Compliance and Intention-to-Treat Analysis

So far, we have assumed that those in the treatment group all get the treatment and those in the control group do not. There are a number of reasons why things are often not as clean as this in practice. Those in the treatment group often cannot be forced to take the de-worming drugs or attend their training program or to take the offered savings package. Similarly, some in the control group may manage to get treatment because they complain or because close substitutes to

the treatment are available outside the experiment. In the presence of imperfect compliance, the probability of receiving treatment among the treatment group is less than one and/or it is more than zero for the control group.

For example, Kling, Liebman, and Katz (2007) provide an evaluation of the Moving To Opportunity (MTO) program in five US cities. This program gave some residents of public housing projects in disadvantaged neighborhoods the opportunity to move out of their public housing. The control group got no new assistance but there were two treatment groups. The S-group received a housing voucher they could use in private rental housing and the E-group the same but with use restricted to areas with poverty rates below 10%. The program is an opportunity to do something, nobody is forced to use the voucher. In fact, only 60% of the S-group and 47% of the E-group did.

The economic interest in this program is the following. It is well-known that cities tend to have residential sorting in which people with similar socioeconomic backgrounds live together. If there are externalities between neighbors then economic theory suggests this sorting may be inefficient. For example, are kids affected by growing up in a bad neighborhood or is their future affected solely by their household characteristics (which tend to be bad in a bad neighborhood)? With non-experimental evidence it has proved very hard to get credible evidence on this issue, but the experimental nature of the MTO program offers a chance to improve our knowledge on this important question.

Let D_i denote actual receipt of the treatment (using the voucher) and let Z_i denote being assigned to the treatment (receiving the voucher). So far we assumed that $Z_i = 1$ implied $D_i = 1$, and $Z_i = 0$ implied $D_i = 0$, but now we depart from this assumption. Individuals with $Z_i = 1$ but D_i are sometimes referred to as ***no-shows***, because they did not show up to get the treatment, and individuals with $Z_i = 0$ but $D_i = 1$ are referred to as ***cross-overs***.

The main concern here is that we are no-longer in a situation of independent treatment, as compliance can be endogenous to potential outcomes. Thus, in general:

$$Y_{1i}, Y_{0i} \not\perp\!\!\!\perp D_i, \tag{14}$$

but, in this case:

$$Y_{1i}, Y_{0i} \perp\!\!\!\perp Z_i. \tag{15}$$

The notation is not casual, as Z_i can be used as an instrumental variable, as discussed in Chapter 4. Alternatively, we can use Z_i as the treatment variable,

instead of D_i:

$$\alpha_{ITT} \equiv \mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0] \tag{16}$$

This parameter is known as ***intention-to-treat*** effect.

B. Longer Run Interaction of Treatment and Intermediate Outcomes

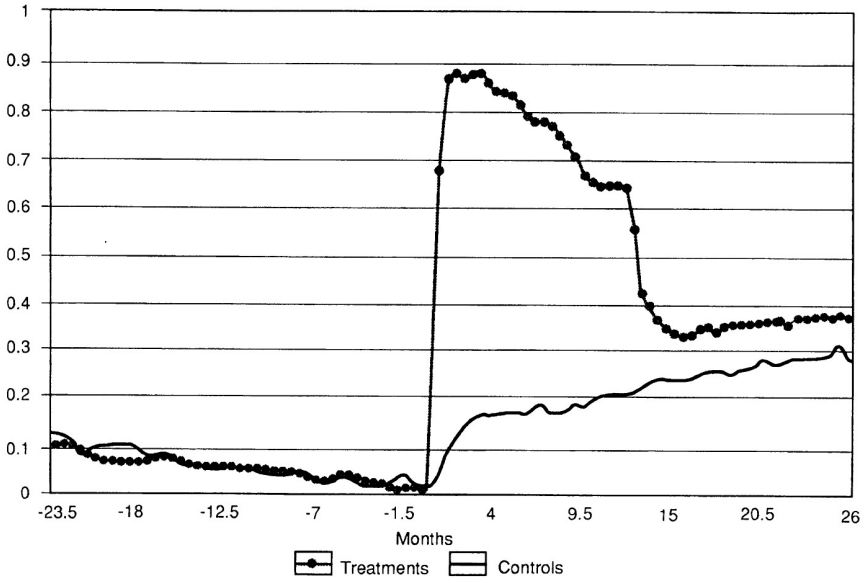
Ham and LaLonde (1996) analyze the National Supported Work program (NSW). The NSW was a training program designed in the U.S. in the mid 1970s to provide training and job opportunities to disadvantaged workers, as part of an experimental demonstration. Ham and LaLonde look at the effects of the NSW on women that volunteered for training. NSW guaranteed to treated participants 12 months of subsidized employment (as trainees) in jobs with gradual increase in work standards. Eligibility requirements were to be unemployed, a long-term AFDC recipient, and have no preschool children. Participants were randomly assigned to treatment and control groups in 1976-1977. The experiment took place in 7 cities. Ham and LaLonde analyze data for 275 women in the treatment group and 266 controls. All volunteered in 1976.

Thanks to randomization, a simple comparison between employment rates of treatments and controls gives an unbiased estimate of the effect of the program on employment at different horizons. Figure 1 below, reproduced from Ham and LaLonde (1996) shows the effects. Initially, by construction there is a mechanical effect from the fact that treated women are offered a subsidized job. As apparent from the figure, compliance with the treatment is decreasing over time, as women can decide to drop from the subsidized job. The employment growth for controls is just a reflection of the program's eligibility criteria. Importantly, after the program ends, a 9 percentage points difference in employment rates is sustained in the medium run, at least until month 26 after the beginning of the program.

But Ham and LaLonde make an important additional point. Even though randomization allows researchers to evaluate the impact of the program on a particular outcome (employment) simply by comparing means, this is not true for any possible outcomes. In particular, if one is interested in the effect of the program on wages or on employment and unemployment durations, a comparison of means would provide a biased estimate of the effect of the program. This is because, as discussed above, the training program had an effect on employment rates of the treated.

To illustrate that, let W_i denote wages, let Y_i be an indicator variable that takes the value of one if the individual is employed, and zero if she is unemployed, and

FIGURE 1. EMPLOYMENT RATES OF AFDC WOMEN IN NSW DEMONSTRATION



Note: This figure corresponds to Figure 1 in Ham and LaLonde (1996)

let η_i denote the ability type, with $\eta_i = 1$ if the individual is skilled, and $\eta_i = 0$ if she is unskilled. Suppose that the treatment increases the employment rates of high skill and low skill workers, but the effect is of less intensity for the high skilled (as they were more likely to find a job anyway without the training program):

$$P(Y_i = 1|D_i = 1, \eta_i = 0) > P(Y_i = 1|D_i = 0, \eta_i = 0), \tag{17}$$

$$P(Y_i = 1|D_i = 1, \eta_i = 1) > P(Y_i = 1|D_i = 0, \eta_i = 1), \tag{18}$$

and:

$$\frac{P(Y_i = 1|D_i = 1, \eta_i = 0)}{P(Y_i = 1|D_i = 0, \eta_i = 0)} > \frac{P(Y_i = 1|D_i = 1, \eta_i = 1)}{P(Y_i = 1|D_i = 0, \eta_i = 1)}. \tag{19}$$

This implies that the frequency of low skill will be greater in the group of employed treatments than in the employed controls:

$$P(\eta_i = 0|Y_i = 1, D_i = 1) > P(\eta_i = 0|Y_i = 1, D_i = 0), \tag{20}$$

which is a way to say that η_i, which is unobserved, is not independent of D_i given $Y_i = 1$, although, unconditionally, $\eta_i \perp\!\!\!\perp D_i$. For this reason, a direct comparison of average wages between treatments and controls will tend to underestimate the effect of treatment on wages. In particular, consider the conditional effects:

$$\Delta_0 \equiv \mathbb{E}[W_i|Y_i = 1, D_i = 1, \eta_i = 0] - \mathbb{E}[W_i|Y_i = 1, D_i = 0, \eta_i = 0], \tag{21}$$

$$\Delta_1 \equiv \mathbb{E}[W_i|Y_i = 1, D_i = 1, \eta_i = 1] - \mathbb{E}[W_i|Y_i = 1, D_i = 0, \eta_i = 1]. \tag{22}$$

Our effect of interest is:

$$\Delta_{ATE} = \Delta_0 P(\eta_i = 0) + \Delta_1 P(\eta_i = 1), \tag{23}$$

whereas the comparison of average wages between treatments and controls gives:

$$\Delta_W = \mathbb{E}[W_i | Y_i = 1, D_i = 1] - \mathbb{E}[W_i | Y_i = 1, D_i = 0]. \tag{24}$$

In general, we shall have $\Delta_W < \Delta_{ATE}$. Indeed, it may not be possible to construct an experiment to measure the effect of training the unemployed on subsequent wages, i.e. it does not seem possible to experimentally undo the conditional correlation between D_i and η_i.