1 Heterogeneous Causal Effects

1.1 Introduction

So far we have estimated regressions of the type:

$$Y_i = \beta_0 + \beta_1 X_i + U_i \tag{1}$$

Here, the parameters of interest are (β_0, β_1) and β_1 corresponds to the causal impact of X on Y under certain assumptions that we have discussed in previous weeks. We have so far assumed that the population parameters β_0 and β_1 are constant across all units in the population, or that they do not change from one individual to another, even though we do not know their exact values. While this assumption of fixed (β_0, β_1) may certainly be valid in many contexts, it may not be so in most contexts. For example, in our classic class size and test score example, it is reasonable to think that different students react differently to being put in classes of different sizes, and in particular that some students are more affected by their class size (for example those that do most of their learning at school due to little resources outside of school) while others may not affected by class size at all (e.g. students with a lot of resources and parental support at home who do most of their learning outside of school).

We refer to the possibility that the causal (or treatment) effect of our independent (policy) variable X on our dependent variable of interest Y varies across individuals as a case of **heterogeneous causal effects** or **heterogeneous treatment effects**. In this case, we consider (β_0, β_1) as random variables that vary from one person to another. The way we deal with heterogeneous treatment effects depends on whether the heterogeneity in treatment effects depends on observable variables in our data. If the heterogeneity does not depend on observables and is instead driven by factors that are unobservable or unknown to the researcher, then the approach we take will depend on whether our regressor X correlates with the error term ϵ or not.

1.2 Case I: Heterogeneity driven by observable factors

If the heterogeneity in treatment effects is driven by factors that are known to us and for which we have data, then we can simply adjust the regression (1) by including interaction terms, as we have already discussed previously. For example, if we are interested in the effect of childbirth on labor force participation of women, then we may posit that the effect will vary for women with a partner who is employed vs. for women with a partner who is not employed. With $birth_i$ being a dummy that takes the value of 1 if woman i gives birth to a child and 0 otherwise, we can consider the following regression:

$$work_i = \beta_0 + \beta_1 birth_i + \beta_2 partner_i + \beta_3 birth_i \times partner_i + u_i$$

Here, $work$ is a dummy that takes the value of 1 if the person is in the labor market and 0 otherwise, and $partner_i$ is also a dummy that takes the value of 1 if the partner of the woman i is employed and 0 otherwise.

The conditional expectations of being in the labor market would be:

$$\begin{aligned}
\mathbb{E}[work|birth = 0, partner = 0] &= \beta_0 \\
\mathbb{E}[work|birth = 1, partner = 0] &= \beta_0 + \beta_1 \\
\mathbb{E}[work|birth = 0, partner = 1] &= \beta_0 + \beta_2 \\
\mathbb{E}[work|birth = 1, partner = 1] &= \beta_0 + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

Here, the coefficients can be interpreted as follows:

β_0 : share of child-bearing women with unemployed partners who work.

β_1 : effect of having a baby on work status.

β_2 : effect of having an employed partner on work status

β_3 : differential effect of having a baby on work status for women with employed partners relative to women with unemployed partners

1.3 Case II: Heterogeneity driven by unobservable factors

While Case I can be sufficient to address heterogeneity, it will be often be the case that our concern is that the heterogeneity in treatment effects is driven by unobservable factors. If this is the case, then the approach that we take will depend on whether our regressor is correlated or not with the unobservable factors (i.e. whether the zero conditional mean assumption holds).

1.3.1 Case IIa: regressor and error term are uncorrelated

Here we assume that the zero conditional mean assumption holds, i.e. $\mathbb{E}(UX) = 0$ in regression (1). For simplicity, we assume that (β_0, β_1) are independent of both U and X.

We know that under the five standard OLS assumptions, the β_{OLS} estimated will be consistent for $\frac{Cov(Y,X)}{Var(X)}$. With (β_0, β_1) being random variables, $\frac{Cov(Y,X)}{Var(X)}$ is no longer equal to β_1[1]. Instead, we will have[2]:

$$
\begin{aligned}
\frac{Cov(Y,X)}{Var(X)} &= \frac{Cov(\beta_0 + \beta_1 X + U, X)}{Var(X)} \\
&= \frac{Cov(\beta_0, X) + Cov(\beta_1 X, X) + Cov(U, X)}{Var(X)} \\
&= \frac{Cov(\beta_1 X, X)}{Var(X)} \text{ since } Cov(\beta_0, X) \text{and } Cov(U, X) \text{ are both } 0 \\
&= \frac{\mathbb{E}(\beta_1 X \cdot X) - \mathbb{E}(\beta_1 X)\mathbb{E}(X)}{Var(X)} \\
&= \frac{\mathbb{E}(\beta_1 X^2) - \mathbb{E}(\beta_1 X)\mathbb{E}(X)}{Var(X)} \\
&= \frac{\mathbb{E}(\beta_1 X(X - \mathbb{E}(X)))}{Var(X)} \\
&= \frac{\mathbb{E}(\mathbb{E}(\beta_1 X(X - \mathbb{E}(X))|X))}{Var(X)} \text{ by the Law of Iterated Expectations} \\
&= \frac{\mathbb{E}(X(X - \mathbb{E}(X))\mathbb{E}(\beta_1|X))}{Var(X)} \text{ since any function of } X \text{ is a constant in the inner conditional expectation} \\
&= \frac{\mathbb{E}(X(X - \mathbb{E}(X))\mathbb{E}(\beta_1))}{Var(X)} \text{ dropping } |X \text{ in the inner expectation since } \beta_1 \text{ is independent of } X \\
&= \mathbb{E}(\beta_1)\frac{\mathbb{E}(X(X - \mathbb{E}(X)))}{Var(X)} \text{ taking out } \mathbb{E}(\beta_1) \text{ out of the expectation since it is a number} \\
&= \mathbb{E}(\beta_1)\frac{\mathbb{E}(X^2 - X\mathbb{E}(X))}{Var(X)} \\
&= \mathbb{E}(\beta_1)\frac{\mathbb{E}(X^2) - \mathbb{E}(X)\mathbb{E}(X)}{Var(X)} \\
&= \mathbb{E}(\beta_1)\frac{Var(X)}{Var(X)} \text{ since } \mathbb{E}(X^2) - \mathbb{E}(X)\mathbb{E}(X) \text{ is the definition of variance} \\
&= \mathbb{E}(\beta_1)
\end{aligned}
$$

So assuming the standard OLS assumptions hold, our β_{OLS} will consistently estimate $\mathbb{E}(\beta_1)$, which is the average value of the treatment effect in the population. Because $\mathbb{E}(\beta_1)$ represents the average value of the random variable β_1 in the population, we refer to it as the **Average Treatment Effect (ATE)**.

[1] The notion "equal to β_1" no longer makes sense since β_1 is now a variable. We can also check this inequality manually.

[2] You do not need to know the following derivation for the exam. This derivation is included in the note for you to see how the end formula is constructed.

1.3.2 Case IIb: regressor and error term are correlated

In this sub-case, we no longer assume that the zero conditional mean assumption holds. Instead, we now assume that our regressor X is endogenous, i.e. $\mathbb{E}(UX) \neq 0$ in regression (1).

As we have seen in previous week, the instrumental variable framework can help us deal with endogenous regressors if we can come up with a valid instrument Z that is both relevant (i.e. correlated with the endogenous variable X) and exogenous (i.e. uncorrelated with unobservable factors U).

Let us assume that we manage to find such Z. By construction, this Z is uncorrelated with U and $\mathbb{E}(UZ) = 0$. Given that the heterogeneity in (β_0, β_1) is also assumed to be driven by the unobservable factors which are uncorrelated with Z, we also consider that (β_0, β_1) are independent of both U and Z.

So suppose that we have

$$Y = \beta_0 + \beta_1 X + U$$

where we are concerned that X is endogenous, and

$$X = \pi_0 + \pi_1 Z + V$$

for some $\pi_1 \neq 0$ and $cov(U, Z) = 0$.

We again assume that (π_0, π_1) are random variables that vary by individual and (π_0, π_1) are independent of (Z, V). We have seen previously that β_{IV} will be consistent for $\frac{Cov(Y,Z)}{Cov(X,Z)}$.

We can simplify $Cov(Y, Z)$ as follows [3]:

$$\begin{aligned}
Cov(Y,Z) &= Cov(\beta_0 + \beta_1 X + U, Z) \\
&= Cov(\beta_0, Z) + Cov(\beta_1 X, Z) + Cov(U, Z) \\
&= Cov(\beta_1 X, Z) \text{ since } Z \text{ is uncorrelated with both } \beta_0 \text{ and } U \\
&= Cov(\beta_1(\pi_0 + \pi_1 Z + V), Z) \\
&= Cov(\beta_1\pi_0, Z) + Cov(\beta_1\pi_1 Z, Z) + Cov(\beta_1 V, Z) \\
&= Cov(\beta_1\pi_1 Z, Z) \text{ since } Z \text{ is uncorrelated with } \beta_1, \pi_0, \pi_1 \text{ and } V \\
&= \mathbb{E}(\beta_1\pi_1 Z \cdot Z) - \mathbb{E}(\beta_1\pi_1 Z)\mathbb{E}(Z) \\
&= \mathbb{E}(\beta_1\pi_1 Z^2) - \mathbb{E}(\beta_1\pi_1 Z)\mathbb{E}(Z) \\
&= \mathbb{E}(\beta_1\pi_1 Z(Z - \mathbb{E}(Z))) \\
&= \mathbb{E}(\mathbb{E}(\beta_1\pi_1 Z(Z - \mathbb{E}(Z))|Z)) \text{ by the Law of Iterated Expectations} \\
&= \mathbb{E}(Z(Z - \mathbb{E}(Z))\mathbb{E}(\beta_1\pi_1|Z)) \text{ since any function of } Z \text{ is a constant in the inner conditional expectation} \\
&= \mathbb{E}(Z(Z - \mathbb{E}(Z))\mathbb{E}(\beta_1\pi_1)) \text{ dropping } |Z \text{ in the inner expectation since } \beta_1 \text{ and } \pi_1 \text{ are independent of } Z \\
&= \mathbb{E}(\beta_1\pi_1)\mathbb{E}(Z(Z - \mathbb{E}(Z))) \text{ taking out } \mathbb{E}(\beta_1\pi_1) \text{ out of the expectation since it is a number} \\
&= \mathbb{E}(\beta_1\pi_1)\mathbb{E}(Z^2 - Z\mathbb{E}(Z)) \\
&= \mathbb{E}(\beta_1\pi_1)(\mathbb{E}(Z^2) - \mathbb{E}(Z)\mathbb{E}(Z)) \\
&= \mathbb{E}(\beta_1\pi_1)Var(Z) \text{since } \mathbb{E}(Z^2) - \mathbb{E}(Z)\mathbb{E}(Z) \text{ is the definition of variance}
\end{aligned}$$

We can also simplify $Cov(X, Z)$ as follows[4]:

[3] You do not need to know the following derivation for the exam. This derivation is included in the note for you to see how the end formula is constructed.

[4] You do not need to know the following derivation for the exam. This derivation is included in the note for you to see how the end formula is constructed.

$$
\begin{aligned}
Cov(X,Z) &= Cov(\pi_0 + \pi_1 Z + V, Z) \\
&= Cov(\pi_0, Z) + Cov(\pi_1 Z, Z) + Cov(V, Z) \\
&= Cov(\pi_1 Z, Z) \text{ since } Z \text{ is uncorrelated with } \pi_0 \text{ and } V \\
&= \mathbb{E}(\pi_1 Z \cdot Z) - \mathbb{E}(\pi_1 Z)\mathbb{E}(Z) \\
&= \mathbb{E}(\pi_1 Z^2) - \mathbb{E}(\pi_1 Z)\mathbb{E}(Z) \\
&= \mathbb{E}(\pi_1 Z(Z - \mathbb{E}(Z))) \\
&= \mathbb{E}(\mathbb{E}(\pi_1 Z(Z - \mathbb{E}(Z))|Z)) \text{ by the Law of Iterated Expectations} \\
&= \mathbb{E}(Z(Z - \mathbb{E}(Z))\mathbb{E}(\pi_1|Z)) \text{ since any function of } Z \text{ is a constant in the inner conditional expectation} \\
&= \mathbb{E}(Z(Z - \mathbb{E}(Z))\mathbb{E}(\pi_1)) \text{ dropping } |Z \text{ in the inner expectation since } \pi_1 \text{ is independent of } Z \\
&= \mathbb{E}(\pi_1)\mathbb{E}(Z(Z - \mathbb{E}(Z))) \text{ taking out } \mathbb{E}(\beta_1\pi_1) \text{ out of the expectation since it is a number} \\
&= \mathbb{E}(\pi_1)\mathbb{E}(Z^2 - Z\mathbb{E}(Z)) \\
&= \mathbb{E}(\pi_1)(\mathbb{E}(Z^2) - \mathbb{E}(Z)\mathbb{E}(Z)) \\
&= \mathbb{E}(\pi_1)Var(Z) \text{ since } \mathbb{E}(Z^2) - \mathbb{E}(Z)\mathbb{E}(Z) \text{ is the definition of variance}
\end{aligned}
$$

With the above simplifications, we can rewrite:

$$
\begin{aligned}
\frac{Cov(Y,Z)}{Cov(X,Z)} &= \frac{\mathbb{E}(\beta_1\pi_1)Var(Z)}{\mathbb{E}(\pi_1)Var(Z)} \\
&= \frac{\mathbb{E}(\beta_1\pi_1)}{\mathbb{E}(\pi_1)} \\
&= \mathbb{E}\left(\frac{\pi_1}{\mathbb{E}(\pi_1)} \cdot \beta_1\right)
\end{aligned}
$$

So under the assumptions made in this case, $\hat{\beta}_{IV}$ that we estimate through two-stage least squares (2SLS) will converge towards $\mathbb{E}(\frac{\pi_1}{\mathbb{E}(\pi_1)} \cdot \beta_1)$. Comparing this formula to the one derived in Case IIa above, we note $\mathbb{E}(\frac{\pi_1}{\mathbb{E}(\pi_1)} \cdot \beta_1)$ is simply a weighted average of $\mathbb{E}(\beta_1)$, where the weights are given by $\frac{\pi_1}{\mathbb{E}(\pi_1)}$ and reflects how individuals are influenced by the instrument. Individuals whose value of X change a lot as their value of Z changes will have relatively high weights while individuals whose value of X change very little as their value of Z changes will receive a low weight. Because the estimation takes into account how *relevant* the instrument is for different individuals, the treatment effects estimated through IV will give us the average treatment effects locally. The term local emphasizes that it is the weighted average that places the most weight on those individuals whose treatment probability is most influenced by the instrument. Thus, $\mathbb{E}(\frac{\pi_1}{\mathbb{E}(\pi_1)} \cdot \beta_1)$ will represent the **Local Average Treatment Effect (LATE)**.

In general, LATE will be different from ATE unless in the following special cases: (i) the treatment effect is the same for all individuals (so β_1 is a fixed constant), (ii) the instrument influences everyone equally (so π_1 is a fixed constant) or (iii) (π_1, β_1) are uncorrelated.

1.4 Summary on heterogeneous treatment effects

The way we deal with this heterogeneity depends on the source of the heterogeneity, and in particular whether such source is observable to us. When these differences depend on observable variables for which we have information, then the heterogeneous causal effects can be estimated using interaction terms in our regressions and we can keep using OLS. However, these differences are very often unobserved (i.e. the reasons why β_1 varies across individuals is unobserved/unknown to us as researchers). In this case, the coefficient estimate from OLS will be consistent for the average value of the treatment effect in the population, or the average treatment effect $\mathbb{E}(\beta_1)$, if the conditional mean

assumption $\mathbb{E}(U|X) = 0$ holds. In the case when causal effects are heterogeneous and our covariate X is endogenous (i.e. $Cov(U, X) \neq 0$), then we can rely on an instrument Z in an instrumental variable framework. With a good instrument Z (i.e. satisfying the exclusion restriction) whose effect is heterogeneous across individuals, IV estimates the local average treatment effect, which assigns weights to individuals based on how likely they react to the instruments. Except for particular cases, LATE will in general be different from ATE.

1.5 True/False Questions

Please respond with true and false to the following questions above.

1. Average Treatment Effect refers to the average value of β_1 in the sample used for the analysis.

2. The weights assigned to individuals to estimate β_{LATE} depend on the heterogeneous effect of the instrument on the endogenous regressor.

3. $\hat{\beta}_{OLS}$ is always consistent for the average treatment effects.

4. LATE will always be lower than ATE since it represents treatment effects for a subset of the population.

5. In an instrumental variable model, LATE will be the same as ATE if the instrument has a homogeneous effect on individuals.

2 Randomized Experiments

Econometricians are mainly concerned with revealing causal relationships, which is actually very difficult and often requires more than a little creativity. Omitted variable bias, in particular, makes this task a hard one. Impact evaluation through experiments seeks to identify the *causal* effect (impact) of an intervention on some selected outcomes, and to quantify these changes in outcomes by running experiments in the real world.

How can we achieve this? The ideal would be to give the policy to certain people, see how they turn out, then go back in time and see what would have happened to those very same people had we not given them the policy. Obviously time travel is not possible. Why can't we just compare these people's outcomes to another group that didn't receive the policy? Well these two groups could be completely different, and we may have difficulties isolating the effect of the program from the other differences between the two groups.

For example, Let's say we want to know the effect of a new health insurance program on health outcomes. Our first thought is to compare the health of those who decide to purchase insurance coverage to the health of those without insurance coverage. What's the problem here? These two populations are likely completely different (think about who seeks out health insurance). How can we be sure that the differences we find among these two groups are solely due to health insurance?

The key issue in measuring impact is to establish a **counterfactual** against which the changes in outcomes for one group induced by the intervention can be measured. The counterfactual should allow

the researcher to *convincingly* measure what would have happened to the beneficiaries in the absence of the intervention.

The techniques we investigate below aim to do just that: they establish a good counterfactual group for those who receive the intervention/program by finding similar individuals that did *not* receive the intervention/program. We refer to the beneficiaries of the intervention or program as the **treatment group** and the non beneficiaries as the **control group**.

2.1 Causal Effects

With randomization, and after verification that there are on average no statistically significant differences between units[5] in the treatment and control groups based on observable characteristics (see below), the measure of impact can be obtained by simple difference in the average outcome between treatment and control groups:

$$Impact = \bar{Y}_T - \bar{Y}_C$$

where $\bar{Y}_T$ is the average outcome for individuals in the treatment group T, and $\bar{Y}_C$ is the average outcome for individuals in the control group C.

In a regression framework, we can retrieve the causal effect by simply regressing our outcome of interest on our treatment variable:

$$Y_i = \beta_0 + \beta_1 T_i + u_i$$

This regression formula works because:

$$\begin{aligned}
\bar{Y}_C &= E[Y_i | i \text{ in Control group}] \\
&= E[\beta_0 + \beta_1 T_i + u_i | i \text{ in Control group}] \\
&= E[\beta_0 | i \text{ in Control group}] + E[\beta_1 T_i | i \text{ in Control group}] + E[u_i | i \text{ in Control group}] \\
&= \beta_0 + 0 + E[u_i | i \text{ in Control group}] \\
&= \beta_0 + E[u_i | i \text{ in Control group}]
\end{aligned}$$

and

$$\begin{aligned}
\bar{Y}_T &= E[Y_i | i \text{ in Treatment group}] \\
&= E[a + \beta_1 T_i + u_i | i \text{ in Treatment group}] \\
&= E[a | i \text{ in Treatment group}] + E[\beta_1 T_i | i \text{ in Treatment group}] + E[u_i | i \text{ in Treatment group}] \\
&= \beta_0 + \beta_1 + E[u_i | i \text{ in Treatment group}] \\
&= \beta_0 + \beta_1 + E[u_i | i \text{ in Treatment group}]
\end{aligned}$$

So, as long as $E[u_i | i \text{ in Treatment group}] = E[u_i | i \text{ in control group}]$

$$\begin{aligned}
\bar{Y}_T - \bar{Y}_C &= \beta_0 + E[u_i | i \text{ in Treatment group}] + \beta_1 - \beta_0 - E[u_i | i \text{ in Control group}] \\
&= \beta_1
\end{aligned}$$

The regression will provide us with an estimate of β_1, as well as standard errors. We can use this to make/test hypotheses about the significance of the treatment variable in explaining our outcome variable of interest.

Because T is a dummy variable indicating treatment, we can interpret the coefficient β_1 as we usually do for a dummy variable: it is an intercept shifter for being in the dummy category compared to the left-out category. In this case, this is exactly the difference in means for Y between the treatment and control groups.

2.2 Key Assumption

The key assumption to recover a causal impact of the treatment is that if it were not for the treatment, the control and the treatment population would be statistically identical, i.e., the two groups have identical

[5] These could be individuals, households, schools, businesses, cities, etc.

expected values for the outcome of interest, regardless of whether they are assigned to treatment/control.

$$E[Y_i|i \text{ in Treatment group}, T] = E[Y_j|j \text{ in Control group}, T]$$

This says that if we hadn't yet administered treatment, we would expect our outcomes (the Y variable) to be the same across both groups.

In a regression framework, the key assumption is the zero conditional mean assumption:

$$E[u_i|T_i = 0] = E[u_i|T_i = 1]$$

In other words, there are no other variables correlated with the outcome Y that are correlated with treatment status. This ensures that $\hat{\beta}_1$ is an unbiased estimator of β_1.

2.3 Tests for Validity of Assumption

We cannot test

$$E[u_i|T_i = 0] = E[u_i|T_i = 1] = 0$$

But we can check to see if the observable characteristics among treatment and control groups are the same on average.

For example, if we want to know the effects of health insurance on health outcomes, and we want to provide some assurance that absent the treatment, our treatment and control groups would have the same outcome, we can show that the average level of education, age, and income (among other variables) are the same across treatment and control groups before the treatment.

What we are doing is testing for "the equality of means of the **observed** characteristics," with the idea that if there is no statistical difference in the observable characteristics (e.g. age, education, income, etc.), then this provides plausible evidence that there is no difference in the **unobserved** characteristics as well (e.g. ability, motivation, level of hypochondria, etc.)

Formally, we want to test that for any observed variable (e.g., age, education, income), the equality

$$E[X_i|i \text{ in Treatment group }] = E[X_i|i \text{ in Control group}]$$

cannot be rejected prior to the treatment being implemented.

2.4 Adding covariates

When we do an RCT, we always want to run a simple regression of the outcome of interest on our treatment variable, as we did above. But we can run additional regressions where we also add observable covariates (additional X / right-hand side variables) to:

1. Add precision to the estimation
2. Verify, as a robustness check, that $\hat{\beta}$ is invariant to the introduction of covariates in the regression

Why don't we expect $\hat{\beta}$ to change? We don't expect our estimate to change precisely because we have randomly selected people to be in the control and treatment group. In other words the observable characteristics of the treatment group shouldn't be correlated with any features of the policy/reform/intervention we gave to the treatment group. But in a given sample, we might by chance observe some statistically significant differences by treatment status for some variables. We can then control for those differences in the regression to again isolate the effect of the treatment.

In a regression framework, we can simply add covariates :

$$Y_i = a + \beta_1 T_i + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

2.5 Heterogeneity

Finally, we can also measure heterogeneity of the program effect for individuals with specific characteristics (such as gender, age, socio-economic status, etc.) by interacting these characteristics with the treatment variable. The implication is that the program has a differential effect on certain subgroups of the population.

In a regression framework, we can simply add an interaction :

$$Y_i = a + \beta_1 T_i + \beta_2 X_{2i} + \beta_3 T_i \times X_{2i} + u_i$$

If the variable X_2 represents a dummy for being female for example, then β_3 gives us the differential effect of the treatment for females relatives to males.

When considering heterogeneous treatment effects, it might make sense to include controls to help contain the risk of omitted variable bias (OVB), caused by potential correlations between the heterogeneous variable and the error term. Here, for example, there could be correlations between the dummy for being female, and the error term, that are not taken into account in the randomization.

2.6 Types of Treatment Estimates

2.6.1 Average Treatment Effect

As we have seen, a really nice thing about randomized controlled trials (RCTs) is that they ensure zero conditional mean assumption, our biggest concern in this class. Sample variation can still lead to some odd results, but this can be controlled for. Ensuring zero conditional mean of our regressor means that we can be pretty confident that the estimate we find for the impact of our treatment (program, intervention, policy, etc.) on a given outcome is *causal*.

For a given treatment T where $T = 1$ for those who received the treatment and $T = 0$ for those who did not, we can estimate the causal effect of the treatment on an outcome Y by regressing

$$Y_i = \beta_0 + \beta_1 T_i + u_i$$

The coefficient β_1 on treatment is referred to as the **Average Treatment Effect (ATE)**. This tells us exactly the difference in the outcome between the treatment and control groups, on average. Whenever we successfully randomize treatment, we can estimate the ATE, and we will almost always be interested in this estimate.

2.6.2 Intention to Treat

Sometimes we don't have perfect **compliance** among the treatment and control groups: some observations in the treatment group do not take the treatment; and some observations in the control group find a way to take the treatment. This decision to "not comply" with the research design is almost certainly correlated with other unobservable characteristics. In this case, if we decide to compare those in the treatment group who took the treatment, to those in the control group who did not, we are no longer comparing the randomly assigned groups, and we won't get a causal estimate.

	Assigned to Treatment	Assigned to Control
Treated	T Complier	Always-Taker
Not Treated	Never-Taker	C Complier

Think of this in the context of a medical randomized control trial. We assign 50 people to control, and 50 people to treatment. The treatment group is given a pill that is supposed to help with energy levels. Among the 50 people in the treatment group, 40 people comply (*T Compliers*) and take the pill but 10 fail to comply and refuse to take the pill (*Never-Takers*). Among the 50 people in the control group, 10 people find a way to get the pill (*Always-Takers*) and 40 people comply and don't take pill (*C Compliers*).

If we were to compare the outcomes of the 40 treatment individuals who took the pill (like they were supposed to), to the 40 control individuals who did not take the pill (like they were supposed to) - then we can no longer say that we are observing the causal effect of the treatment. This is precisely because certain people "selected" into these treatment and control groups based on unobservables we may or may not see, and hence can't always control for. In our case, maybe those who didn't comply in the treatment group by not taking the pill are generally people with greater energy levels. Similarly maybe those in the control group who took the pill are hypochondriacs and are always complaining of illnesses.

What do we do in this case? We acknowledge that we don't have perfect compliance, but we continue to "naively" compare those who were *assigned* to the treatment group, to those who were *assigned* to the control group. So if we are in the treatment group but didn't comply,we will still be considered treated, and if we are in the control group and didn't comply, we will still be considered control. In keeping with the initial allocation of treatment, we call the estimator "**Intention to Treat**" (ITT). We estimate the ITT using the same specification as the ATE - the only difference is that we acknowledge that our treatment assignment was not perfectly successful, so we no longer estimate the ATE.

We can always estimate the ITT with an RCT. If the treatment only affects the outcome through its effect in varying a characteristic of interest (taking a pill in the example above), the ITT tells us the effect of that characteristic on the outcome, diluted by the fact that not every treated observation complied with the treatment (e.g., took the pill).

Formally, in the example above with pills to increase energy levels, suppose we have a measure of individual energy levels (y) and we also know rates of take-up of the pill (P) by treatment status (T or C). Then we have

$$\bar{Y}_T = P_T * E[Y|P = 1, T = 1] + (1 - P_T) * E[Y|P = 0, T = 1]$$
$$\bar{Y}_C = P_C * E[Y|P = 1, T = 0] + (1 - P_C) * E[Y|P = 0, T = 0]$$

Here, P_T is the probability that a unit in the treatment group takes the pill and is given to be 0.8. Similarly, P_C is the probability that a unit in the control group takes the pill and is given to be 0.2.

We need to also consider the breakdown of compliers, never-takers, and always-takers. Since 20% of control individuals took the pill and treatment was randomized, we assume that 20% of the population overall are always-takers. Similarly, since 20% of treatment individuals didn't take the pill, we assume 20% of the population are never-takers. Then we can write:

$$\bar{Y}_T = 0.2 * E[Y|Always - Taker, T = 1] + 0.6 * E[Y|Complier, T = 1] + 0.2 * E[Y|Never - Taker, T = 1]$$
$$\bar{Y}_C = 0.2 * E[Y|Always - Taker, T = 0] + 0.6 * E[Y|Complier, T = 0] + 0.2 * E[Y|Never - Taker, T = 0]$$

Noting that randomization of treatment means expected outcomes should be the same for always-takers and never-takers (since these groups differ only in their treatment assignment but not in their treatment take-up), we end up with

$$\begin{aligned}\bar{Y}_T - \bar{Y}_C &= 0.6 * (E[Y|Complier, T = 1] - E[Y|Complier, T = 0]) \\ &= E[Complier] * (E[Y_T|Complier] - E[Y_C|Complier])\end{aligned}$$

The ITT estimate thus gives us the impact of the treatment among compliers, weighted by the share of compliers. Note that if everyone complies with the treatment, then the ITT is the same as the ATE.

2.6.3 Treatment on the Treated

When compliance with random assignment is not perfect, how close the estimate of the ITT is to the ATE will depend on a few things, most importantly what the take-up (share of treatment compliers) is (and any differences in treatment effects).

In the example above with pills to increase energy levels, policymakers may care most about the impact of taking the pill on energy levels and not worry about the fact that there is selection into who takes

the pill. The reasoning for this could be that the population that always comply are the population of interest here. In this case, policymakers would want to simply estimate the average impact of the treatment among compliers, which we refer to as "**Treatment on the Treated**" (TOT) estimate. TOT is given by the following formula:

$$E[Y|Complier, T=1] - E[Y|Complier, T=0] = E[Y_T|Complier] - E[Y_C|Complier] = (\bar{Y}_T - \bar{Y}_C)/E[Complier]$$

where, $\bar{Y}_T$ and $\bar{Y}_C$ are as defined in the previous subsection above.

Note that the TOT will always be larger than the ITT, since it is not diluted by the null impact of the treatment on individuals assigned to treatment that did not actually get treated.

If the characteristics of compliers are very different from those of the full population, the TOT may look very different from the ATE. For example, if the treatment individuals who took the energy pill were those who thought they could benefit most from a boost in their energy levels, the TOT might be greater than the ATE. The only situation when the TOT will equal the ATE is if compliers experience the same treatment effects as the population would on average, i.e., if

$$E[Y_T|Complier] - E[Y_C|Complier] = E[Y_T] - E[Y_C]$$

So anytime we don't have perfect compliance with treatment, we need to be concerned about heterogeneity in treatment effects and how these could create differences between the TOT and ATE.

2.7 Exercise

Policymakers want to study the impact of schooling on wages in the labor market. To generate random variation in years of education, these policymakers run a program that pays parents if their children stay in school. Households are divided into two groups. The design randomly *encourages* a first group of households to increase their children's years of education, without varying years of education directly. To the second group of households, no encouragement design is given and they can decide to keep their children at school or have drop out of school. Children in these two groups of households are then followed for 10 years and their work status in the labor market outcomes is tracked.

1. Do you think that the encouragement design is a good instrument for years of school? Please justify your answer

2. Assume that there is a perfect compliance of the intervention (i.e. the instrument is equally effective for all households in the treatment group), would ITT and ATE be the same or be different here?

3. Assume that there is a 50% compliance with the encouragement design (i.e. the instrument is only effective for 50% of the households). Can you express ITT as a function of ATE? Can you express TOT as a function of ITT?

4. ***Only for this part.*** Consider the regression of years of education on the encouragement design, and assume that the context of the intervention is such that boys regularly attend schools regardless of incentives while girls often drop out of school early. As a researcher, you observe the gender composition of the children of households, can you propose a regression model that accounts for this heterogeneity in the effect of the encouragement design on the years of education?