Chapter 1. Potential Outcomes and Causality: Treatment Effects

Joan Llull

Quantitative & Statistical Methods II
Master in Economics of Public Policy
Barcelona School of Economics

I. Introduction

According to the definition in *The New Palgrave: A Dictionary of Economics*, econometrics is the branch of economics that aims to give empirical content to economic relations. We use econometrics to this end in three different ways:

- ***Inference of association*** (or descriptive statistics): to quantitatively describe or summarize the co-movement of different economic variables. These are widely used to establish "facts" that motivate economic theory, but do not aim to provide, by themselves, inference about the mechanisms beyond the economic relation between the variables. They have recently regained popularity with the recent availability of huge datasets (big data).
- ***Forecasting***: to predict one variable based on the movements of other variables, without any aim, once again, of describing the economic mechanisms of interest. For example, the presence of many people carrying an umbrella can help us in predicting that it is likely to rain, but no-one would think that it is going to rain *because* many people carry an umbrella.
- ***Causal inference***: the process of drawing a conclusion about a cause-effect connection between economic variables. For example, we may be interested in whether hospital stays improve the health of inpatients. The association between hospital stay and health would clearly be negative, as inpatients are usually sicker than average. However, is it because hospitals worsen health of inpatients (e.g. exposition to other illnesses) or simply because unhealthy individuals go to hospitals? Causal inference aims to answer *what if* type of questions. Hence, we want to know whether *the same inpatient* would be more or less healthy if she was sent to home versus being kept at the hospital. Thus, we want to net out the fact that individuals that go to hospitals are unhealthier to begin with.

In this course we focus on the latter, which is a powerful tool for policy evaluation.

The evaluation of public (and private) policies is very important for efficiency, and ultimately to improve welfare. There is a vast literature in economics, mostly

in public economics, but also in development economics and labor economics, devoted to the evaluation of different programs. Examples include training programs, welfare programs, wage subsidies, minimum wage laws, taxation, Medicaid and other health policies, school policies, feeding programs, microcredit, and a variety of other forms of development assistance. These analyses aim at quantifying the effects of these policies on different outcomes, and ultimately on welfare.

The classic approach to quantitative policy evaluation is the ***structural approach***. This approach specifies a class of theory-based models of individual choice, chooses the one within the class that best fits the data, and uses it to evaluate policies through simulation. This approach has the main advantage that it allows both *ex-ante* and *ex-post* policy evaluation, and that it permits evaluating different variations of a similar policy without need to change the structure of the model or reestimate it (out of sample simulation). The main critique to this approach, though, is that there is a host of untestable functional form assumptions that undermine the force of the structural evidence because they have unknown implications for the results, give researchers too much discretion, and its complexity often affects transparency and replicability. Some people has argued that this approach puts too much emphasis on external validity at the expense of a more basic internal validity.

During the last two decades, the ***treatment effect approach*** has established itself as an important competitor that has introduced a different language, different priorities, techniques, and practices in applied work. This approach has changed the perception of evidence-based economics among economists, public opinion, and policy makers. The main goal of this approach is to evaluate (*ex-post*) the impact of an existing policy by comparing the distribution of a chosen outcome variable for individuals affected by the policy (the treatment group), with the distribution of unaffected individuals (control group). The main challenge of this approach is to find a way to perform the comparison in such a way that the distribution of outcome for the control group serves as a good counterfactual for the distribution of the outcome for the treated group in the absence of treatment. The main focus of this approach is in the understanding of the sources of variation in data with the objective of identifying the policy parameters, even though these parameters are formally not valid representations of the outcomes of implementing the same policy in an alternative environment, or of implementing variations of the policy even to the same environment. Thus, this approach helps in the assessment of future policies in a more informal way.

The main advantage of this approach is that, given its focus on internal validity,

the exercise gives transparent and credible identification. The main disadvantage is that estimated parameters are not useful for welfare analysis because they are not deep parameters (they are reduced-forms instead), and as a result, they are not policy-invariant. In that respect, a treatment effect exercise is less ambitious.

In order to set up a treatment effects analysis (or essentially any econometric analysis that aims at making causal inference) we have to formulate essentially four crucial questions:

- What is the causal relation of interest?

- What experiment could ideally be used to capture the causal effect of interest? Some times this would be trivial. For example, if we are interested in the causal effect of a subsidy on education, we just have to give the subsidy to some people and keep some otherwise equal individuals without any subsidy, and the compare the outcomes of both groups. However, if we are interested in the effect of being black on wages (everything else equal), it is more difficult to think of an ideal experiment, as we cannot transform a group of blacks into whites. If our interest is on hiring probabilities or starting wage, one could send some fake curriculum vitae of black and white workers with the same characteristics and compare the rate at which calls are returned. Other cases are even more difficult (so much that we say that they are fundamentally unidentified). For example, if we are interested in the effect of age of start of school on test scores, we have the following problem. On the one hand, if we compare individuals born in different months (so that they are in different grades but of similar age), we have the problem that those who started earlier have accumulated more schooling at the time of the test. If, instead, we take the test at the end of the first (or second) grade, we have the problem that the ones in the lower grade are more mature at the time of the test. Thus, it is fundamentally not possible to identify the effect of early schooling on early test scores (for later outcomes the maturity differences vanish somehow).

- What is your "identification strategy" (in the terminology of Angrist and Krueger, 1999)? This is, how do you use observational data to approximate a real experiment?

- What is your method for inference? Or, in other words, the population to be studied, the sample to be used, and the assumptions made when constructing standard errors.

II. Potential Outcomes, Selection Bias, and Treatment Effects

Consider the population of individuals that are susceptible of a treatment. Let Y_{1i} denote the outcome for an individual i if exposed to the treatment ($D_i = 1$), and let Y_{0i} be the outcome for the same individual if not exposed ($D_i = 0$). The ***treatment effect*** for individual i is thus $Y_{1i} - Y_{0i}$. Note that Y_{1i} and Y_{0i} are ***potential outcomes*** in the sense that we only observe one of the two:

$$Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i). \tag{1}$$

This poses the main challenge of this approach, as the treatment effect can not be computed for a given individual. Fortunately, our interest is not in treatment effects for specific individuals *per se*, but, instead, in some characteristics of their distribution, like some average.

We mainly focus on two parameters of interest. The first one is the ***average treatment effect*** (ATE):

$$\alpha_{ATE} \equiv \mathbb{E}[Y_{1i} - Y_{0i}], \tag{2}$$

and the second one is ***average treatment effect on the treated*** (TT):

$$\alpha_{TT} \equiv \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]. \tag{3}$$

Note the subtle difference between the two. The first one is an ideal parameter of interest, but difficult to obtain: the average of the treatment effects for the population. The second one, which is often easier to obtain, is the average treatment effect computed over treated individuals, that is, for the individuals that actually experiment treatment.

The reason why the second parameter is easier to identify is, precisely, that we only observe Y_i. Let β denote the difference in mean outcomes for treated and untreated individuals, which can be rewritten as:

$$\begin{aligned}\beta &\equiv \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\ &= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]}_{\alpha_{TT}} + \underbrace{(\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0])}_{\text{selection bias}}.\end{aligned} \tag{4}$$

The second term, which we call "selection bias" indicate the difference in untreated potential outcomes between treated and untreated individuals. A nonzero bias may result from a situation in which treatment status is the result of individual decisions where those with low Y_0 choose treatment more frequently than those with high Y_0 or vice versa. In the hospital example, treated individuals (those

hospitalized) would also be less healthy had they been at home (negative bias). Thus, the comparison of health between inpatients and other individuals gives a negatively biased estimate of the effect of hospitalization.

From a structural model of D_i and Y_i, one could obtain the implied average treatment effects. Here, they are instead defined with respect to the distribution of potential outcomes, so that, relative to the structure, they are ***reduced-form causal effects***. Econometrics has conventionally distinguished between reduced form effects, uninterpretable but useful for prediction, and structural effects, associated with rules of behavior. The treatment effects provide this intermediate category between predictive and structural effects, in the sense that recovered parameters are causal effects, but, as reduced form effects, they are uninterpretable outside of the sample/population of interest and the treatment implemented (or, in other words, they lack external validity). Furthermore, an important assumption of the potential outcome representation is that the effect of the treatment on one individual is independent of the treatment received by other individuals. This excludes equilibrium or feedback effects, as well as strategic interactions among agents. Hence, the framework is not well suited to the evaluation of system-wide reforms which are intended to have substantial equilibrium effects.

Sample analogs for α_{ATE} and α_{TT} are:

$$\alpha^S_{ATE} \equiv \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}}(Y_{1i} - Y_{0i}) \tag{5}$$

$$\alpha^S_{TT} \equiv \frac{1}{N_1}\sum_{\mathrm{i}=1}^{\mathrm{N}} D_i(Y_{1i} - Y_{0i}), \tag{6}$$

where $N_1 \equiv \sum_{\mathrm{i}=1}^{\mathrm{N}} D_i$ is the number of treated individuals. If factual and counterfactual potential outcomes were observed, these quantities could be estimated without error. However, since they are not, the distinction is not very useful on practical grounds. Importantly, though, depending on whether we estimate population (α) or sample (α^S) average treatment effects, standard errors will be different, so we should take this into account when computing confidence intervals. The sample average version of β is given by:

$$\begin{aligned}\beta^S &\equiv \bar{Y}_T - \bar{Y}_C \\ &\equiv \frac{1}{N_1}\sum_{\mathrm{i}=1}^{\mathrm{N}} Y_i D_i - \frac{1}{N_0}\sum_{\mathrm{i}=1}^{\mathrm{N}}(1 - D_i)Y_i,\end{aligned} \tag{7}$$

where $N_0 \equiv N - N_1$ is the number of untreated (or control) individuals.

III. Identification of Treatment Effects under Different Assumptions

The identification of the treatment effects depends on the assumptions we make on the relation between potential outcomes and the treatment. The easiest case is when the distribution of the potential outcomes is independent of the treatment:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i. \tag{8}$$

This situation is typical in randomized experiments, where individuals are assigned to treatment or control in a random manner. For example, this occurs, for a given school, in the random assignment of pupils to different class sizes implemented in a randomized experiment called STAR that we will discuss as an example in next chapter. When this is the case, $F(Y_{1i}|D_i = 1) = F(Y_{1i})$, and $F(Y_{0i}|D_i = 0) = F(Y_{0i})$, which implies that $\mathbb{E}[Y_{1i}] = \mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_i|D_i = 1]$ and $\mathbb{E}[Y_{0i}] = \mathbb{E}[Y_{0i}|D_i = 0] = \mathbb{E}[Y_i|D_i = 0]$, and, as a result, $\alpha_{ATE} = \alpha_{TT} = \beta$. Thus, an unbiased estimate of α_{ATE} is given by the difference between average outcomes of treated and control individuals:

$$\widehat{\alpha}_{ATE} = \bar{Y}_T - \bar{Y}_C = \beta^S. \tag{9}$$

In this context, there is no need to "control" for other covariates, unless there is direct interest in their marginal effects, or we want to compute effects for specific groups (we return to this point below).

A less restrictive assumption is ***conditional independence***:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i|X_i, \tag{10}$$

where X_i is a vector of covariates. This situation is known as matching, as for each "type" of individual (i.e. each value of covariates) we can match treated and control individuals, so that the latter act as counterfactuals for the former. Conditional independence implies that the above results are valid for a given X_i, that is $\mathbb{E}[Y_{1i}|X_i] = \mathbb{E}[Y_{1i}|D_i = 1, X_i] = \mathbb{E}[Y_i|D_i = 1, X_i]$ and $\mathbb{E}[Y_{0i}|X_i] = \mathbb{E}[Y_{0i}|D_i = 0, X_i] = \mathbb{E}[Y_i|D_i = 0, X_i]$, and, as a result:

$$\begin{aligned}
\alpha_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}] &= \mathbb{E}[\mathbb{E}[Y_{1i} - Y_{0i}|X_i]] \\
&= \int \mathbb{E}[Y_{1i} - Y_{0i}|X_i] dF(X_i) \\
&= \int (\mathbb{E}[Y_i|D_i = 1, X_i] - \mathbb{E}[Y_i|D_i = 0, X_i]) dF(X_i).
\end{aligned} \tag{11}$$

In words, the bottom expression computes the difference in average observed outcomes of treated and control individuals that share each value of X_i, and integrate

over the distribution of X_i. Thus, it is "matching" treated individuals with controls that share the same X_i. Similarly, the treatment effect on the treated is:

$$\begin{aligned}\alpha_{TT} &= \int \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1, X_i]dF(X_i|D_i = 1) \\ &= \int \mathbb{E}[Y_i - \mathbb{E}[Y_{0i}|D_i = 1, X_i]|D_i = 1, X_i]dF(X_i|D_i = 1) \\ &= \int \mathbb{E}[Y_i - \mu_0(X_i)|D_i = 1, X_i]dF(X_i|D_i = 1), \end{aligned} \tag{12}$$

where $\mu_0(X_i) \equiv \mathbb{E}[Y_i|D_i = 0, X_i]$, and we use the fact that $\mathbb{E}[Y_i|D_i = 0, X_i] = \mathbb{E}[Y_{0i}|X_i] = \mathbb{E}[Y_{0i}|D_i = 1, X_i]$. The function $\mu_0(X_i)$ is used as an imputation device (matching) for Y_{0i}.

Finally, sometimes we cannot assume conditional independence:

$$(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp D_i|X_i. \tag{13}$$

In this case, we will need some variable Z_i that provides ***exogenous variation*** in the treatment, meaning that it satisfies the independence assumption:

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i|X_i, \tag{14}$$

and the relevance condition:

$$Z_i \not\perp\!\!\!\perp D_i|X_i. \tag{15}$$

As we discuss in Chapter 4, in this context we are only going to be able to identify an average treatment effect for a subgroup of individuals, and we call the resulting parameter a ***local average treatment effect***.

IV. Linear Regression and Treatment Effects

The potential outcomes notation is very useful to think about causality, but it can be cumbersome. Rearranging the terms in Equation (1) yields:

$$\begin{aligned} Y_i &= Y_{0i}(1 - D_i) + Y_{1i}D_i \\ &= Y_{0i} + (Y_{1i} - Y_{0i})D_i \\ &= \mathbb{E}[Y_{0i}] + (Y_{1i} - Y_{0i})D_i + (Y_{0i} - \mathbb{E}[Y_{0i}]) \\ &\equiv \beta_0 + \beta_i D_i + U_i. \end{aligned} \tag{16}$$

Equation (16) gives an expression for the causal effect of D_i on $_i$, which is given by the random coefficient β_i.

Note that β_i is different for different individuals. To fix ideas, assume initially that $\beta_i = \beta$ for all i. In this case, β is the coefficient of a linear regression of outcome on treatment dummy. Since β_i is a constant, $Y_{1i} - Y_{0i}$ is also a constant, and $\alpha_{ATE} = \alpha_{TT} = \beta$. As in a linear regression, consistently estimating β requires that the error term U_i, or equivalently, Y_{0i}, is independent the treatment variable D_i.

If β_i is not a constant, we can also obtain the average treatment effect on the treated α_{TT} by a similar linear regression. Consider the following linear regression:

$$Y_i = \beta_0 + \bar{\beta} D_i + U_i. \tag{17}$$

Recall from undergraduate econometrics that the slope coefficient in a linear regression with constant is the ratio between the covariance of the outcome and the regressor divided by the variance of the regressor. Thus:

$$\bar{\beta} = \frac{\mathrm{Cov}(Y_i, D_i)}{\mathrm{Var}(D_i)} = \frac{\mathbb{E}[Y_i D_i] - \mathbb{E}[Y_i]\,\mathbb{E}[D_i]}{\mathbb{E}[D_i^2] - \mathbb{E}[D_i]^2}. \tag{18}$$

To operate this expression, we first note that D_i only takes values of zero and one, and, thus, $D_i^2 = D_i$. Given this, the denominator boils down to:

$$\mathbb{E}[D_i^2] - \mathbb{E}[D_i]^2 = \mathbb{E}[D_i](1 - \mathbb{E}[D_i]). \tag{19}$$

To operate the numerator, we appeal to the law of iterated expectations and the fact that $\Pr(D_i = 1) = \mathbb{E}[D_i]$. Given this, the first term of the covariance is:

$$\begin{aligned} \mathbb{E}[Y_i D_i] &= \mathbb{E}[Y_i \cdot 1 | D_i = 1] \Pr(D_i = 1) + \mathbb{E}[Y_i \cdot 0 | D_i = 0] \Pr(D_i = 0) \\ &= (\beta_0 + \mathbb{E}[\beta_i | D_i = 1] + \mathbb{E}[U_i | D_i = 1])\,\mathbb{E}[D_i], \end{aligned} \tag{20}$$

the first element of the second term is:

$$\begin{aligned} \mathbb{E}[Y_i] &= (\beta_0 + \mathbb{E}[\beta_i | D_i = 1] + \mathbb{E}[U_i | D_i = 1])\,\mathbb{E}[D_i] + (\beta_0 + \mathbb{E}[U_i | D_i = 0])(1 - \mathbb{E}[D_i]) \\ &= \beta_0 + \mathbb{E}[\beta_i | D_i = 1]\,\mathbb{E}[D_i] + \mathbb{E}[U_i | D_i = 1]\,\mathbb{E}[D_i] + \mathbb{E}[U_i | D_i = 0](1 - \mathbb{E}[D_i]), \end{aligned} \tag{21}$$

and the covariance is:

$$\begin{aligned} &\mathbb{E}[Y_i D_i] - \mathbb{E}[Y_i]\,\mathbb{E}[D_i] \\ &\quad = \{\mathbb{E}[\beta_i | D_i = 1] + (\mathbb{E}[U_i | D_i = 1] - \mathbb{E}[U_i | D_i = 0])\}\,\mathbb{E}[D_i](1 - \mathbb{E}[D_i]). \end{aligned} \tag{22}$$

Thus, the regression coefficient is:

$$\bar{\beta} = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1] + (\mathbb{E}[Y_{0i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0]) = \beta, \tag{23}$$

which is equal to the average treatment effect on the treated plus the selection bias.

To finish with this chapter, we expand a bit on the connection between regression and treatment effects. In particular, we reestablish the notion of conditional independence in the regression context, and introduce a few extra discussions associated to it. These points are illustrated with an example about the effects of schooling on wages.

A. Conditional independence

Consider the treatment effect analysis that studies the effect of education on wages. Let C_i denote the treatment, such that $C_i = 1$ if individual i goes to college, and $C_i = 0$ otherwise. Let Y_{1i} denote the earnings of this individual if she attends college, and Y_{0i} her earnings if she does not. A regression of observed wages on a college dummy provides an estimate of β, as discussed above. In this context, however, it is plausible that the selection bias is not zero. In particular, individuals with more "ability" are more likely to obtain education and also, they are more productive in the labor market, whether they get education or not. Thus, the sample of individuals with $C_i = 1$ has, on average, higher ability than those with $C_i = 0$, and, thus, their wages when they do not study (and also those when they study) are, on average, higher. In other words, this regression exaggerates the benefits of college, as:

$$\mathbb{E}[Y_{0i}|C_i = 1] - \mathbb{E}[Y_{0i}|C_i = 0] > 0, \tag{24}$$

and, thus $\beta > \alpha_{TT}$.

However, if we can "control" for ability, A_i, the independence assumption is more plausible:

$$\mathbb{E}[Y_{0i}|A_i, C_i = 1] - \mathbb{E}[Y_{0i}|A_i, C_i = 0] = 0. \tag{25}$$

In words, this means that, for a given level of ability A_i, individuals that go to college are not systematically different than those who do not go.

B. Omitted variable bias

In terms of a regression, we typically think of including A_i as a control in the regression:

$$Y_i = \beta_0 + \beta C_i + \gamma A_i + U_i, \tag{26}$$

even though we could additionally add an interaction term between A_i and C_i. Under the conditional independence assumption, we think that the long regression

in Equation (26) has a causal interpretation, whereas the short one in (17) has not.

The ***omitted variable bias formula*** provides a connection between the parameters identified in the two equations. Note the connection between the two equations by defining $\tilde{U}_i \equiv \gamma A_i + U_i$, and rewritting (26) as:

$$Y_i = \beta_0 + \beta C_i + \tilde{U}_i. \tag{27}$$

Now, the regression coefficient, denoted by $\tilde{\beta}$, equals:

$$\begin{aligned}\tilde{\beta} &= \frac{\text{Cov}(Y_i C_i)}{\text{Var}(C_i)} \\ &= \frac{\text{Cov}(\beta_0 + \beta C_i + \tilde{U}_i), C_i}{\text{Var}(C_i)} \\ &= \beta + \frac{\text{Cov}(\tilde{U}_i, C_i)}{\text{Var}(C_i)}.\end{aligned} \tag{28}$$

The second term is the omitted variable bias. We can rewrite the omitted variable bias formula as:

$$\begin{aligned}\frac{\text{Cov}(\tilde{U}_i, C_i)}{\text{Var}(C_i)} &= \frac{\text{Cov}(\gamma A_i + U_i, C_i)}{\text{Var}(C_i)} \\ &= \gamma \frac{\text{Cov}(A_i C_i)}{\text{Var}(C_i)} + \frac{\text{Cov}(U_i C_i)}{C_i} \\ &= \gamma \frac{\text{Cov}(A_i C_i)}{\text{Var}(C_i)}.\end{aligned} \tag{29}$$

We establish that the second term in the central expression is equal to zero using the conditional independence assumption, which implies $\mathbb{E}[U_i|A_i, C_i] = 0$, the law of iterated expectations, and that $\mathbb{E}[U_i] = \mathbb{E}[Y_{0i} - \mathbb{E}[Y_{0i}]] = 0$ by construction:

$$\text{Cov}(U_i, C_i) = \mathbb{E}[U_i C_i] - \mathbb{E}[U_i]\,\mathbb{E}[C_i] = \mathbb{E}[U_i C_i] = \mathbb{E}[\mathbb{E}[U_i|A_i, C_i]C_i] = 0. \tag{30}$$

Thus, the omitted variable bias, given by Equation (29) equals to γ, which determines the mapping between ability and outcome, and $\text{Cov}(A_i C_i)/\,\text{Var}(C_i)$, which is the slope coefficient of a regression that related ability and the treatment. In our example, the treatment and ability are positively associated, and ability is also likely to be positively related to outcomes, so the omitted variable bias would be positive (we overestimate the effect of schooling on wages).

To illustrate this, we reproduce below Table 3.2.1 from Angrist and Pischke, which estimates such regressions with data from the National Longitudinal Survey of Youth (1979). The advantage of this dataset is that contains information about the Armed Forces Qualification Test (AFQT), which can be used to control

for ability. We also introduce age dummies, which can be associated with experience, some other controls, including mother's and father's education and dummy variables for race and census region, and occupation dummies, even though this might be a bad idea as discussed below.

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age Dummies	Col. (2) and Additional Controls*	Col. (3) and AFQT score	Col. (2) with Occupation Dummies
	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)

Note: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2,434.
* Additional controls are mother's and father's education and dummies for race and census region.

Results from the table suggest that ability is an important omitted covariate, and that treatment is not independent of ability. In particular, the coefficient in Column (3) is already somewhat smaller than those in Columns (1) and (2), and the coefficient in Column (4) is substantially smaller. This is so because mother's and father's education can be correlated with children's ability, and AFQT is positively correlated with ability. Since ability is positively correlated with schooling and wages, the omitted variable bias is positive.

There are two final remarks to make about the results in the previous table. First, note that years of schooling (which is the treatment variable) is not a zero-one variable in this case. Second, we haven't discussed Column (5). The remaining of the chapter addresses these two points.

C. Treatment variables that take more than two values

Even though we discuss continuous treatments in more detail below this example motivates that we discuss now discrete treatments with more than two possible values, like schooling. In this case, we define the treatment effect function as:

$$Y_{si} \equiv f_i(s). \tag{31}$$

In our college vs high school example, $Y_{1i} = f_i(16)$ and $Y_{0i} = f_i(12)$. In other words, $f_i(s)$ answers the causal "what if" question for every possible value of s, even if the observed schooling is a particular value S_i. This treatment effect function can be linear (i.e. $f_i(s) - f_i(s-1) = f_i(s') - f_i(s'-1)$ for any $s \neq s'$) or not.

D. Endogenous controls

Finally, we discuss the introduction of occupation as a control. Occupation is could be endogenous to the treatment, as college educated individuals are more likely to work in white collar occupations than high school dropouts. Thus, define the observed white collar dummy as:

$$W_i = W_{1i}C_i + W_{0i}(1 - C_i), \tag{32}$$

where $W_i = \mathbb{1}\{\text{white collar}\}$. For simplicity, assume that C_i is randomly allocated across individuals. Thus, we can estimate the average treatment effects:

$$\mathbb{E}[Y_i|C_i = 1] - \mathbb{E}[Y_i|C_i = 0] = \mathbb{E}[Y_{1i} - Y_{0i}], \tag{33}$$

and:

$$\mathbb{E}[W_i|C_i = 1] - \mathbb{E}[W_i|C_i = 0] = \mathbb{E}[W_{1i} - W_{0i}]. \tag{34}$$

However, if we estimate the causal effect of college on wages conditional on occupation (for example, conditional on $W_i = 1$) we obtain:

$$\mathbb{E}[Y_i|W_i = 1, C_i = 1] - \mathbb{E}[Y_i|W_i = 1, C_i = 0] = \mathbb{E}[Y_{1i}|W_{1i} = 1] - \mathbb{E}[Y_{0i}|W_{0i} = 1]. \tag{35}$$

Thus, we are subtracting apples and oranges, as the population of individuals with $W_{1i} = 1$ is different from that with $W_{0i} = 1$. The first group is individuals who work in white collar if they go to college (but not necessarily do so if they only complete high school), whereas the second group is the set of individuals that work in white collar when they do not go to college (and probably also do so if they go). Specifically, we can rewrite Equation (35) to obtain:

$$\begin{aligned}\mathbb{E}[Y_i|W_{1i} = 1] - \mathbb{E}[Y_i|W_{0i} = 1] \\ = \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|W_{1i} = 1]}_{\text{causal effect}} + \underbrace{(\mathbb{E}[Y_{0i}|W_{1i} = 1] - \mathbb{E}[Y_{0i}|W_{0i} = 1])}_{\text{selection bias}}.\end{aligned} \tag{36}$$

In the example of returns to years of education, the coefficient falls from 0.087 to 0.066. However, it is hard to say what we should make of this decline. The selection bias in this context can be positive or negative, depending on the relation between occupational choice, school attendance, and potential earnings. This change in the coefficient may simply be an artifact of the selection bias. So we would do better to control only for variables that are not themselves caused by education.