

1 Example: Wages and Education

Suppose we are interested in studying the “causal effect” of college education on wages. In other words, we want to know whether college *causes* a change in wages, and by how much.

How do we operationalize the idea of a causal effect? One way, familiar from economic modeling, is to try to obtain a *ceteris paribus* effect. That is, we want to imagine holding all other factors *constant*.

Recall that in the classical regression model, when we write down

$$E[y_i|x_i] = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},$$

we can interpret, say, β_2 as the “effect” of a one unit increase in x_{i2} on the expected value of y_i , holding fixed x_{i3}, \dots, x_{ik} . So, if we want to interpret β_2 as the “causal” effect of x_{i2} on the outcome, we need to have a good set of additional regressor variables.

Let’s work through a very simple example to get some insights. Suppose that y_i is hourly wage of an individual, and

$$s_i = \begin{cases} 0 & \text{high school graduate} \\ 1 & \text{college graduate} \end{cases}$$

Assume that everyone in our population of study is either a HS graduate or a college graduate.

Suppose

$$E[y_i|s_i] = \beta_1 + \beta_2 s_i.$$

(This is not restrictive, because s_i is binary, so there are only two possible conditional means.) So

$$\beta_2 = E[y_i|s_i = 1] - E[y_i|s_i = 0].$$

As you showed in a homework, the OLS estimator for β_2 is

$$\hat{\beta}_2 = \frac{1}{n_1} \sum_i s_i y_i - \frac{1}{n_0} \sum_i (1 - s_i) y_i,$$

where n_1 is the number of observations with $s_i = 1$, and n_0 is the number of observations with $s_i = 0$. Notice that $s_i y_i$ is zero if $s_i = 0$, and is equal to y_i if $s_i = 1$.

We might be concerned that β_2 does not have a causal interpretation, because college graduates might be different in other ways (than just educational attainment) than high school graduates. For example, children who go to college might have had a different home environment (on average) than those who did not go to college, and this home environment might also be related to wages.

Suppose we have an additional variable, parent's education p_i , equal to 1 if one or more parents went to college, and 0 otherwise. We might view this as a partial measure of home environment.

To fix ideas, suppose that s_i, p_i have the following joint distribution in the study population:

$$\begin{aligned} Pr(p_i = 0, s_i = 0) &= .4 \\ Pr(p_i = 0, s_i = 1) &= .1 \\ Pr(p_i = 1, s_i = 0) &= .1 \\ Pr(p_i = 1, s_i = 1) &= .4 \end{aligned}$$

So children tend to get similar educational levels as their parents. Also, note that

$$Pr(p_i = 0 | s_i = 0) = \frac{.4}{.4 + .1} = \frac{4}{5}, \quad Pr(p_i = 1 | s_i = 0) = \frac{.1}{.4 + .1} = \frac{1}{5},$$

etc.

Also, suppose that

$$E[y_i | s_i, p_i] = \gamma_1 + \gamma_2 s_i + \gamma_3 p_i.$$

(I am using gammas instead of betas, because we will want to compare this specification to the earlier one in a moment.)

Using the law of iterated expectations and the assumptions about the joint distribution of p_i, s_i , we can write

$$\begin{aligned} E[y_i | s_i = 0] &= E[E[y_i | s_i = 0, p_i]] \\ &= E[y_i | s_i = 0, p_i = 0]Pr(p_i = 0 | s_i = 0) + E[y_i | s_i = 0, p_i = 1]Pr(p_i = 1 | s_i = 0) \\ &= \gamma_1(4/5) + (\gamma_1 + \gamma_3)(1/5) \\ &= \gamma_1 + \gamma_3/5. \end{aligned}$$

By similar reasoning,

$$E[y_i | s_i = 1] = \gamma_1 + \gamma_2 + \gamma_3(4/5).$$

So,

$$E[y_i | s_i = 1] - E[y_i | s_i = 0] = \gamma_2 + \gamma_3(3/5).$$

We can also write

$$E[y_i | s_i] = (\gamma_1 + \gamma_3/5) + (\gamma_2 + \gamma_3(3/5))s_i.$$

This is equivalent to our original model, setting

$$\begin{aligned} \beta_1 &= \gamma_1 + \gamma_3/5 \\ \beta_2 &= \gamma_2 + \gamma_3(3/5). \end{aligned}$$

Note that:

- **Both models are correct.** They just refer to different conditioning variables.

- The simple differences-in-means estimator $\hat{\beta}_2$ will be unbiased for β_2 , since the model $E[y_i|s_i] = \beta_1 + \beta_2 s_i$ is correct. However, it will not be unbiased for γ_2 . In particular, if $\gamma_3 > 0$ (which seems plausible), then $\hat{\beta}_2$ will tend to overestimate γ_2 .
- If we don't observe p_i , we could still determine the sign of the bias of $\hat{\beta}_2$ for γ_2 if we know the sign of γ_3 and can make some assumptions about the joint distribution of s_i and p_i .

2 Omitted Variables and OLS

We want to generalize the example above to situations where variables are omitted from a least squares analysis. Suppose that we have IID data and that a classical regression model holds:

$$E[y_i|x_i] = x_i' \gamma,$$

where x_i is $k \times 1$ and γ is $k \times 1$.

Let x_{i1} and x_{i2} denote subvectors of x_i with dimension k_1 and k_2 , respectively, so that $x_i' = (x_{i1}', x_{i2}')$. Write

$$E[y_i|x_i] = x_{i1}' \gamma_1 + x_{i2}' \gamma_2.$$

If we condition only on x_{i1} , then

$$E[y_i|x_{i1}] = x_{i1}' \gamma_1 + E[x_{i2}|x_{i1}]' \gamma_2.$$

In general, $E[x_{i2}|x_{i1}]$ is a function of x_{i1} . If it is a linear function, say $E[x_{i2}|x_{i1}] = D' x_{i1}$, then

$$E[y_i|x_{i1}] = x_{i1}' \gamma_1 + x_{i1}' D \gamma_2 = x_{i1}' (\gamma_1 + D \gamma_2).$$

Then we have a linear model

$$E[y_i|x_{i1}] = x_{i1}' \beta,$$

where

$$\beta = (\gamma_1 + D \gamma_2) \neq \gamma_1.$$

If we run a least squares regression of y on x_1 then the OLS coefficients will be unbiased for β , but biased for γ_1 unless $D \gamma_2 = 0$.

If instead $E[x_{i2}|x_{i1}]$ is a nonlinear function of x_{i1} , then $E[y_i|x_{i1}]$ will no longer have a linear conditional mean. Instead, the conditional mean will be

$$E[y_i|x_{i1}] = x_{i1}' \gamma_1 + d(x_{i1}),$$

where $d(x) = E[x_{i2}|x_{i1} = x] \gamma_2$.

In any case the conditional mean will not be equal to $x_{i1}' \gamma_1$, so the least squares coefficients will not be unbiased or consistent for γ_1 . What does OLS converge to in this case? Basically, it estimates a linear approximation to $E[y_i|x_{i1}]$. Recall that OLS would solve

$$\min_{\beta} \sum_{i=1}^n (y_i - x_{i1}' \beta)^2.$$

The solution would not change if we scaled the minimand by $1/n$:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^n (y_i - x'_{i1}\beta)^2.$$

Consider the following minimization problem:

$$\min_{\pi} E[(y_i - x'_{i1}\pi)^2].$$

To actually solve this problem, we would need to know the joint distribution of (y_i, x_{i1}) .

Now, suppose the conditional mean of y_i given x_{i1} is linear: $E[y_i|x_{i1}] = x'_{i1}\beta$. Then it can be shown that the solution is $\pi = \beta$.

If instead the conditional mean is nonlinear, then there is still a solution for π , which is sometimes called the “best linear predictor” coefficient. It gives the best linear prediction for y_i given x_{i1} .

Notice the similarity between the best linear predictor problem and the OLS minimization problem. If the sample size is large, we expect that

$$\frac{1}{n} \sum_i (y_i - x'_{i1}\beta)^2 \approx E[(y_i - x'_{i1}\beta)^2].$$

So the OLS coefficient will converge in probability to π , the best linear predictor coefficient.