Multiple Regression: Motivation

Suppose you are interested in understanding how wages are related to years of education, so you look at the model

$$wage = \beta_1 + \beta_2 educ + v.$$

For now, think of v as being the typical disturbance term.

Now ask yourself: are there any other variables that are correlated with both education and wage? I am strongly inclined to say "yes." Take IQ for example; I would expect someone with a high IQ to receive more education than average, but to also receive a higher wage than average *even without more education* by virtue of having a high IQ.

Okay, but how does this affect our analysis? That we fail to include a variable that is correlated with both the independent and dependent variable means our estimate for β_2 will be **biased**, that is, $E[b_2] \neq \beta_2$. We refer to this as **omitted variable bias**.

To see why, consider someone who has one more year of education. An additional year of education is correlated with a higher wage. But more education is also correlated with a higher IQ, which itself is correlated with a higher wage. Because we have omitted IQ from our model, we are unintentionally attributing the effect of higher IQ to the effect of education. In other words, we are failing to hold IQ constant when considering different levels of education – we are getting both the effect of higher education *and* the effect of higher IQ in our estimate of β_2.

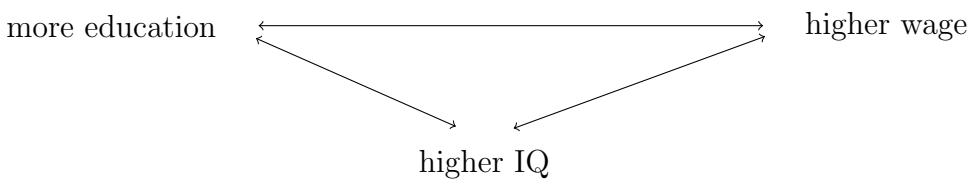

FIGURE 1: More education is correlated with higher wage, but it's also correlated with higher IQ. If we do not hold IQ constant, then we are not accurately characterizing the relationship between education and wage.

So how do we progress? Simple: just stick IQ into the regression as well. Our improved model is thus of the form

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + u.$$

Now when we take the partial derivative with respect to education, we are explicitly holding

IQ constant by definition of a partial derivative. Therefore

$$\frac{\partial wage}{\partial education} = \beta_2$$

gives the relationship between education and wage where IQ is being *controlled for*.

Of course, there are probably other omitted variables as well. In a laboratory experiment, all of these factors can be controlled for if the experiment is properly designed. But we are limited to the data we observe, which may or may not contain all relevant variables. (Probably won't.) Thus, even if we control for a bunch of variables, we still can never fully assert a causal relationship based on nothing but observational data.

Example: Wages

Import `wages.csv` into R. It contains, you guessed it, information about (monthly) wages, education, IQ, and some other stuff. If we regress wages on education, the result is

$$\widehat{wage} = 146.952 + 60.214 educ.$$

This implies that someone with one more year of education would be expected to have a higher monthly wage by \$60.214. But as discussed earlier, this is implicitly including the effect of a higher IQ, since the model above fails to control for IQ. We control for IQ by regressing wage on both education and IQ. By doing so, we expect the effect of education to be lower because now the effect isn't being exaggerated by a higher IQ. Indeed,

$$\widehat{wage} = -128.890 + 42.058 educ + 5.138 IQ.$$

The relevant R code is found at the end.

Dummy Variables

We might be interested in seeing how different categories affect the dependent variable. For instance, we might want to see if someone working in an urban environment earns more than someone working elsewhere. To analyze, we construct a **dummy variable** that is equal to either zero or one. An urban worker would have value $urban = 1$, and a non-urban worker would have value $urban = 0$. Accordingly, we would run the regression

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 urban + u.$$

The coefficient β_4, then, would tell you the expected difference in monthly wage for an urban worker compared to a non-urban worker.

Dummy Variable Trap

Notice above that there are two categories, but only one dummy variable. In general, if you have n categories, then you must include exactly $n-1$ dummy variables; the category you omit is called the **reference category**. Including dummy variables for all possible categories results in the **dummy variable trap**, which is a source of **perfect multicollinearity** that breaks OLS estimation (explained below). So always use one fewer dummy than there are categories.

Here's a really stupid example to illustrate why things go wrong. Suppose everyone has a choice of either having Swedish Fish, Sour Patch Kids, or Mike and Ikes, but can only choose one. We want to see how many cavities each person receives from eating so much damn candy. We record their choices in the following manner:

$$\begin{aligned} choice = 1 \quad & \text{if Swedish Fish,} \\ choice = 2 \quad & \text{if Sour Patch Kids,} \\ choice = 3 \quad & \text{if Mike and Ikes.} \end{aligned}$$

Now let's create dummies for all categories. Let $d_1 = 1$ for choosing Swedish Fish; $d_2 = 1$ for choosing Sour Patch Kids; and $d_3 = 1$ for choosing Mike and Ike. Then the possible values for each dummy are

$$\begin{aligned} choice = 1 \quad &\implies \quad d_1 = 1,\ d_2 = 0,\ d_3 = 0, \\ choice = 2 \quad &\implies \quad d_1 = 0,\ d_2 = 1,\ d_3 = 0, \\ choice = 3 \quad &\implies \quad d_1 = 0,\ d_2 = 0,\ d_3 = 1. \end{aligned}$$

Notice that in all three cases, $d_1 + d_2 + d_3 = 1$. And therefore, say, $d_1 = 1 - d_2 - d_3$. This is perfect multicollinearity because one of our regressors (d_1) can be perfectly explained by a linear relationship of two other regressors (d_2 and d_3). So if we try to regress *cavities* on d_1, d_2, and d_3, then OLS explodes and we're all doomed.

Except you can just remove any one of the three dummies from the regression, then all is well and well is all for all. The coefficients of the model are then seen as being *relative* to the reference category. To that end, consider the model where we omit the Swedish Fish

dummy variable d_1, given by

$$cavities = \beta_1 + \beta_2 d_2 + \beta_3 d_3 + u.$$

Let us interpret each coefficient.

- β_1: how many cavities are associated with eating Swedish Fish (the reference category);
- β_2: how many more (or less, if negative) cavities are associated with eating Sour Patch Kids instead of Swedish Fish;
- β_3: how many more (or less, if negative) cavities are associated with eating Mike and Ikes instead of Swedish Fish.

Example: Wages

Again using `wages.csv`, let us consider the regression proposed earlier,

$$wage = \beta_1 + \beta_2 educ + \beta_3 IQ + \beta_4 urban + u.$$

OLS estimation yields

$$\widehat{wage} = -209.926 + 39.767 educ + 5.127 IQ + 157.466 urban.$$

The result is statistically significant, so we conclude that an urban worker is expected to earn a monthly wage \$157.466 higher than that of a non-urban worker.

R Code

```
1  library("stargazer")
2  wages <- read.csv("wages.csv")
3
4  reg1 <- lm(wage ~ educ, data = wages)
5  stargazer(reg1, type = "text")
6
7  reg2 <- lm(wage ~ educ + IQ, data = wages)
8  stargazer(reg2, type = "text")
9
10 reg3 <- lm(wage ~ educ + IQ + urban, data = wages)
11 stargazer(reg3, type = "text")
```