1 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 error 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**. Technically this is consequence of violating classical OLS assumption 2 (see below), i.e. zero conditional mean, because $E[v|educ] \neq 0$.

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.

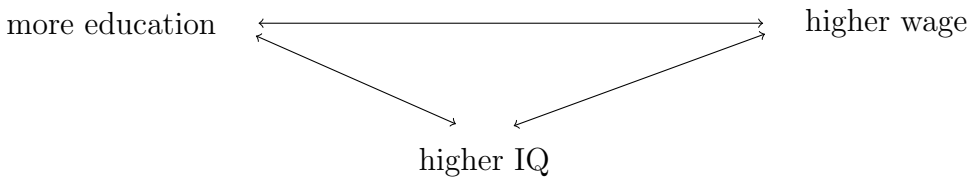

FIGURE 1: More education is correlated with higher wage, but it's also correlated with higher IQ. If we do not hold IQ constant, then we are not accurately characterizing the relationship between education and wage.

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.

2 Classical OLS Assumptions

For OLS to "work" by default, we need the following conditions to hold.

1. *Correct Linear Model.* The true model is linear and correctly specified as

   $$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u.$$

   Intuition: if we estimate the wrong model, then our results are probably garbage.

2. *Zero Conditional Mean.* The error term has zero mean conditional upon the regressors,

   $$E[u|x_2, \ldots, x_k] = 0.$$

   Intuition: think of the error term as being the, well, error of the model. If we expect the error to be non-zero, then we essentially expect our model to be mistaken, on average. Which means our model is probably garbage.

   More technically, it allows us to go from

   $$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u$$

   to

   $$E[y|x_2, \ldots, x_k] = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k,$$

   the latter being our equation for fitted values and predictions of y.

3. *Homoskedasticity.* The conditional variance of the error term is constant and finite,

$$\text{Var}(u|x_2, \ldots, x_k) = \sigma_u^2 < \infty.$$

There isn't much economic intuition here; it's mostly a technical assumption, albeit an unrealistic one, that offers a convenient starting point for rigorous analysis. In practice it is violated frequently, which is not difficult to deal with.

4. *Uncorrelated Errors.* For any two errors u_i and u_j such that $i \neq j$,

$$\text{Cov}(u_i, u_j) = 0.$$

Intuition: if model errors are correlated, then there is some underlying pattern that we are overlooking, so our results are probably garbage.

Suppose we look at ECN 102 final exam scores in all of 2017; that means we're looking at ECN 102 final exam scores for three different professors. Problem is, different professors write exams of differing difficulty. Hence we would expect a lenient professor's students to do better than the regression predicts (so we'd have correlation among observations with positive u), and we expect a challenging professor's students to do worse than what the regression predicts (so we'd have correlation among observations with negative u). This is called **clustering** because each final exam forms a cluster of students. Note that we have correlated errors *within* a cluster; but not *across* clusters.

5. *Normality of Errors.* Error terms are normally distributed with some variance σ^2,

$$u_i \sim \mathcal{N}\left(0, \sigma^2\right).$$

This is another technical assumption for "nice" results. In practice it can be weakened.

6. *No Perfect Multicollinearity.* There exists no exact linear relationship between explanatory samples. Furthermore, the number of observations must be greater than the number of explanatory variables (plus constant term), i.e. $n \geq k$.

Intuition: if there is such a perfect relationship between two or more regressors, then we can't "untangle" the effect of each regressor. In other words, it's like including the same regressor twice, and that redundancy breaks the OLS solution technique.

Assumptions 1-2 imply that OLS estimates are unbiased, so that $E[b_j] = \beta_j$. Assumptions 1-4 imply that OLS estimates are consistent, so that $b_j \xrightarrow{p} \beta_j$ as $n \to \infty$. Furthermore,

assumptions 1-4 imply that OLS is the **b**est **l**inear **u**nbiased **e**stimator, or BLUE. When we say "best," we mean we have the smallest standard errors and hence precision of inference is the most accurate. If we throw in assumption 5, then OLS is the **b**est **u**nbiased **e**stimator, even when compared to nonlinear methods. (Note that assumption 5 is needed to do inference with small samples; this should not be too surprising since we've already had to assume normality for small sample inference repeatedly in this course.)

3 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.

3.1 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. 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

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

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.