1 Multiple Regression

1.1 Motivation: Omitted Variables

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. The interpretation is that we want to explain *wage* with *educ* and "other stuff" captured in v.

Now ask yourself: of the "other stuff" in v that explains wage, is any of that also correlated with education? I am strongly inclined to say "yes." Take IQ for example. I would expect a higher IQ to explain a higher wage; but I also suspect that there is a correlation between IQ and years of education (e.g. college students have a higher IQ than the general public). So when we consider someone with more education, we are also likely considering someone with a higher IQ. This is problematic because β_2 in the regression above is implicitly telling us the effect of education *and* of IQ on wage, and therefore β_2 does not isolate the effect of education on wage.

In other words, *we are failing to hold IQ constant when considering different levels of education*, and consequently we are getting both the effect of higher education *and* the effect of higher IQ in our estimate of β_2. This relationship is illustrated in Figure 1.

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 $E[v|educ] \neq 0$.

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 exper-

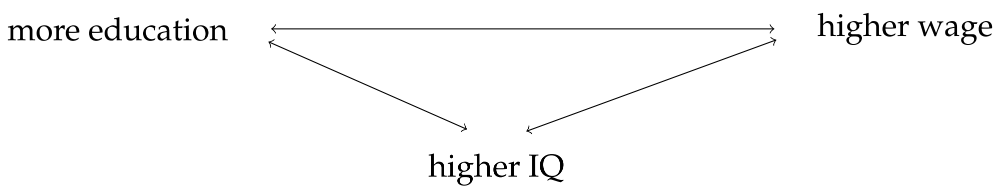

FIGURE 1: More education is correlated with higher wage, but it's also correlated with higher IQ. If we do not hold IQ constant, then we are not accurately characterizing the relationship between education and wage.

iment, ideally 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 be certain that we have fully determined the direct relationship between any x and y.

1.2 Example: Wages

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

$$\widehat{wage} = 139.12 + 61.59 \times educ.$$

This implies that someone with one more year of education would be expected to have a higher monthly wage by \$61.586. 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,

$$\widehat{wage} = -131.67 + 44.27 educ + 4.95 IQ.$$

So as predicted, the estimated effect of education on wage drops from 61.59 to 44.27. Before controlling for IQ, our estimate of β_2 had an *upward bias*.

The relevant Stata commands and output are found on the following page.

2 Classical OLS Assumptions

For OLS to "work" by default, we need the following conditions to hold given dependent variable y and the set of regressors $x_2, x_3, \ldots, x_k$. Note that we have $k-1$ regressors

```
. regress wage educ

      Source |       SS           df       MS      Number of obs   =       852
-------------+----------------------------------   F(1, 850)       =    107.82
       Model |  15622714.1         1  15622714.1   Prob > F        =    0.0000
    Residual |   123165191       850  144900.225   R-squared       =    0.1126
-------------+----------------------------------   Adj R-squared   =    0.1115
       Total |   138787905       851   163088.02   Root MSE        =    380.66

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   61.58627   5.931167    10.38   0.000     49.94482    73.22772
       _cons |   139.1171   81.16017     1.71   0.087    -20.18069     298.415
------------------------------------------------------------------------------

. regress wage educ iq

      Source |       SS           df       MS      Number of obs   =       852
-------------+----------------------------------   F(2, 849)       =     67.17
       Model |  18960227.2         2  9480113.62   Prob > F        =    0.0000
    Residual |   119827678       849  141139.786   R-squared       =    0.1366
-------------+----------------------------------   Adj R-squared   =    0.1346
       Total |   138787905       851   163088.02   Root MSE        =    375.69

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   44.26802   6.851945     6.46   0.000     30.81928    57.71676
          iq |   4.954005   1.018755     4.86   0.000     2.954432    6.953578
       _cons |  -131.6712    97.5547    -1.35   0.177    -323.1479    59.80543
------------------------------------------------------------------------------
```

because we started at x_2. Therefore we will be estimating k things because we are also estimating the intercept coefficient. Hence we will have $n-k$ degrees of freedom when we do inference.

1. **MLR1: Correct Linear Model.** The true model is linear and correctly specified as

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u. \tag{1}$$

   Intuition: if we estimate a population model that's actually of a different form, then our estimates are probably garbage.

2. **MLR2: Zero Conditional Mean.** The error term has zero mean conditional upon the regressors, that is,

$$E[u|x_2, \ldots, x_k] = 0. \tag{2}$$

   Intuition: think of the error term as being the mistake of the model. If we expect the mistake to be non-zero on average, then our model is probably garbage. This condition is equivalent to saying that u is uncorrelated with all of the regressors.

More technically, it allows us to go from

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

$$E[y|x_2, \ldots, x_k] = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k, \tag{3}$$

the latter being the interpretation of the regression line itself (i.e. the conditional expectation of y given our regressors).

3. **MLR3: Homoskedasticity.** The conditional variance of the error term is constant and finite, that is,

$$\text{Var}(u|x_2, \ldots, x_k) = \sigma_u^2 < \infty. \tag{4}$$

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 (as explained later). This condition is illustrated in Figure 2.

4. **MLR4: Independent Errors.** Errors for different observations are statistically independent, that is,

$$u_i \perp u_j \quad \text{whenever } i \neq j.$$

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

As an example of a violation, suppose we look at ECN 102 final exam scores in all of

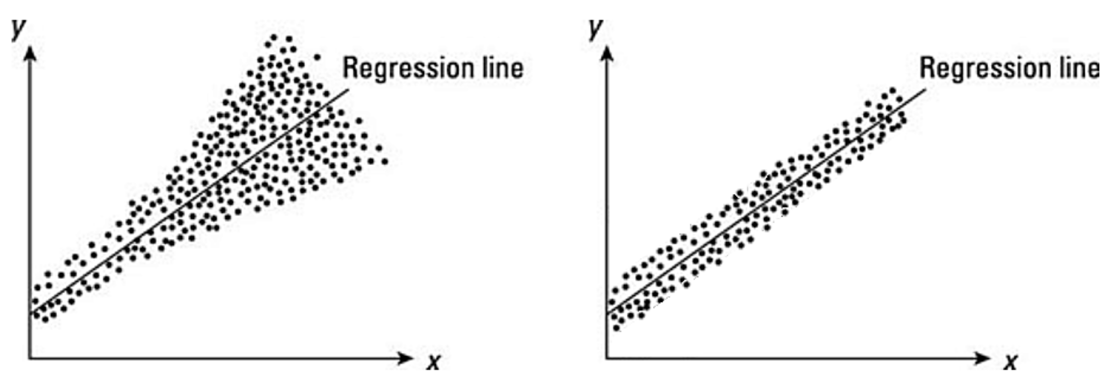

FIGURE 2: The figure on the left is an example of heteroskedasticity; the right an example of homoskedasticity. The left is heteroskedastic because the variation around the regression line gets bigger as x increases. Good luck envisioning this in higher dimensions.

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 professor's final exam forms a cluster of students. (Note that students in different clusters, however, are independent from each other.)

5. **MLR5: Normality of Errors.** Errors are normally distributed with some variance σ^2, that is,

$$u_i \sim \mathcal{N}\left(0, \sigma^2\right). \tag{5}$$

   This is another technical assumption for "nice" results, explained below. In practice it can be weakened, but it is necessary for inference on small sample sizes.

6. **MLR6: 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.

3 Implications of OLS Assumptions

You can see that most of these assumptions are close analogues to the simple regression, the exception being MLR6. You will not be surprised then to learn that the implications are largely the same as well.

- Assumptions MLR1-2 imply that OLS estimates are unbiased, so that $E[b_j] = \beta_j$.
- Assumptions MLR1-4 imply that OLS estimates are consistent, so that $b_j \xrightarrow{p} \beta_j$ as $n \to \infty$. Furthermore, assumptions MLR1-4 imply that OLS is the **b**est **l**inear **u**nbiased **e**stimator, or **BLUE**. When we say "best," we mean they have the smallest standard errors and hence precision of inference is the most accurate.

- Adding MLR5 implies that OLS is the **b**est **u**nbiased **e**stimator, or **BUE**, even when compared to nonlinear methods. Furthermore, it implies that

$$t \equiv \frac{b_j - \beta_j}{\text{se}(b_j)} \sim T(n-k)$$

  is exactly true for any β_j, even for small samples; without MLR5 it is only approximately true if the sample size is large enough. (Therefore MLR5 is required for inference on small samples.) We are estimating k things, which is why we have $n-k$ degrees of freedom.

- Assumption MLR6 is absolutely required; in the presence of perfect multicollinearity, the regression cannot be executed. Accordingly, this is usually just implicitly assumed because otherwise it's game over and we should just give up and go home. (Actually, there's usually a very easy fix for it, shown in a bit.)

4 Multiple Regression Inference

Under MLR1-4, the t-statistic regarding regressor x_j is given by

$$t = \frac{b_j - \beta_j}{\text{se}(b_j)}, \tag{6}$$

and it is drawn from an approximate $T(n-k)$ distribution.Inference proceeds in the usual way. There is no rule of thumb for how large n needs to be for the approximation to be adequate. If MLR5 holds, then t is drawn from exact $T(n-k)$ distribution.

If either MLR3 or MLR4 fail, then the typical standard errors are not valid. We can oftentimes use one of the following alternatives, however.

- `vce(robust)`: use **heteroskedasticity-robust standard errors** if only MLR3 fails
- `vce(cluster, x)`: use **cluster-robust standard errors** if MLR4 alone fails because of suspected clustering in variable x
- use **heteroskedasticity and autocorrelation-consistent (HAC) standard errors** if using time series data.

Autocorrelation means the value of a variable today is highly correlated with its value in previous periods (e.g. GDP data). Time series is a different animal in Stata, so don't worry about the command for it.

5 Dummy Variables

5.1 Definition of Dummy Variable

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. Another way of thinking about it is, β_4 captures the expected change in wage if a worker moves from a non-urban environment to an urban environment, that is, if *urban* changes from 0 to 1.

5.2 Dummy Variable Trap

Notice in the preceding example that there are two categories, but only one dummy variable. In general, if you have m categories, then you must include exactly $m - 1$ dummy variables; the category you omit is called the **reference category**. Including dummy variables for all 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 silly example to illustrate why things go wrong. People become really loyal to stupid things that don't matter, for example, which brand of cola they drink.[1] They either drink Coke and only Coke; or Pepsi and only Pepsi; or, for the purposes of this example, RC Cola and only RC Cola.[2]

We want to see how many cavities people receive from drinking a beverage that is sometimes used to remove rust from nails. We record their preference in the following manner:

$$choice = 1 \text{ if Coke}, \quad choice = 2 \text{ if Pepsi}, \quad choice = 3 \text{ if RC Cola}.$$

[1]Blind taste test? People can't tell the difference.
[2]No one actually drinks RC Cola, do they?

Now let's create dummies for all categories. Let $d_1 = 1$ for choosing Coke; $d_2 = 1$ for choosing Pepsi; and $d_3 = 1$ for choosing RC Cola. 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 Coke 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 being a Coke drinker (reference category);
- β_2: how many more (or less, if negative) cavities are associated with being a Pepsi drinker instead of a Coke drinker;
- β_3: how many more (or less, if negative) cavities are associated with being an RC Cola drinker instead of a Coke drinker.

In the case of the urban workers, β_4 captures how much higher of a wage a person receives if they work in an urban environment relative to working in a non-urban environment (the reference category).

5.3 Example: Wages

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

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

OLS estimation yields

$$\widehat{wage} = -213.28 + 41.58educ + 4.92IQ + 169.01urban.$$

As shown in the Stata output below, the p-value for β_4 indicates statistically significance, so we conclude that an urban worker is expected to earn a monthly wage \$169.01 higher than that of a non-urban worker. To account for the possibility of heteroskedasticity, I tell Stata to use heteroskedasticity-robust standard errors with the option `vce(robust)`.

```
. regress wage educ iq urban, vce(robust)

Linear regression                               Number of obs     =        852
                                                F(3, 848)         =      53.41
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1719
                                                Root MSE          =     368.15

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   41.58144   6.793912     6.12   0.000     28.24658    54.91629
          iq |   4.919558    .944874     5.21   0.000     3.064992    6.774124
       urban |   169.0137   26.54763     6.37   0.000      116.907    221.1205
       _cons |  -213.2816   95.91454    -2.22   0.026    -401.5393   -25.02381
------------------------------------------------------------------------------
```

6 Interactions

6.1 Marginal Effects

When we have multiple regressors, we might be interested in how they, um, interact with each other when it comes to explaining the dependent variable.

A regression with interactions will look something like

$$y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 xz + u,$$

where xz is the interaction term. The idea is that x might affect y differently depending on what value z is, and vice versa. That is, the marginal effect of x on y is given by

$$\frac{dy}{dx} = \beta_2 + \beta_4 z.$$

6.2 Example: Foreign Aid and Dictatorships

You might be interested in how foreign aid affects education funding in undeveloped countries, so you run the regression

$$educ = \beta_1 + \beta_2 aid + u.$$

The coefficient β_2 tells you the association between an additional dollar of foreign aid received and education funding for the average undeveloped country; the marginal effect is just a constant. That is, one more dollar of foreign aid is associated with β_2 more dollars of education funding.

We suspect, however, that the effect of foreign aid is different depending whether the undeveloped country is democratic or ruled by a dictator. Introduce the dummy variable $dictator = 0$ for democracy and $dictator = 1$ for dictatorship and run the regression

$$educ = \beta_1 + \beta_2 aid + \beta_3(aid \times dictator) + u.$$

In this formulation, the effect of foreign aid depends on the value of *dictator* (i.e. the interaction of regressors). The marginal effect of foreign aid on education funding is

$$\frac{\partial educ}{\partial aid} = \beta_2 + \beta_3 \times dictator.$$

If the country is a democracy, then the marginal effect of foreign aid on education funding is just β_2. If the country is a dictatorship, then the marginal effect is $\beta_2 + \beta_3$. A natural hypothesis is that $\beta_3 < 0$, or in words: dictatorships that receive foreign aid don't seem to allocate as much of that foreign aid into education when compared to a democracy. (A more nuanced approach would try to measure the degree of dictatorship instead of a binary designation, but you get the picture.)