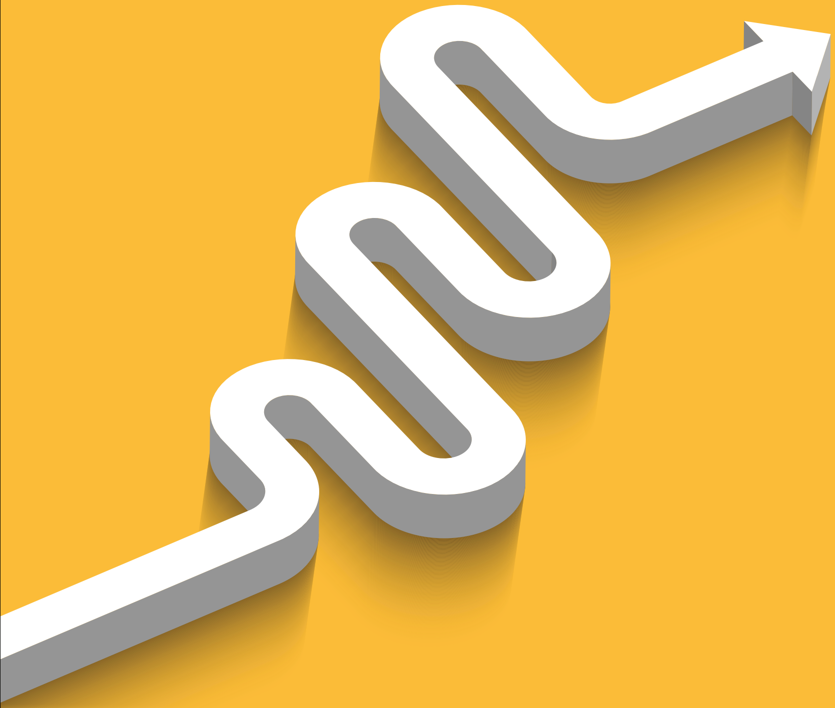
GLOBAL EDITION

A Practical Guide to Using Econometrics



SEVENTH EDITION

A. H. Studenmund

Pearson

β

USING ECONOMETRICS

Table 5.2 *(continued)*

Country	P	GDPN	CV	N	CVN	PP	PC	DPC
Austria	139.53	69.6	1.24	3.52	35.2	0	0	0
Netherlands	137.29	75.2	1.54	6.40	24.1	1	0	0
Belgium	101.73	77.7	3.49	4.59	76.0	1	0	1
France	91.56	81.9	25.14	24.70	101.8	1	0	1
Luxembourg	100.27	82.0	0.10	0.17	60.5	1	0	1
Denmark	157.56	82.4	0.70	2.35	29.5	1	0	0
Germany, West	152.52	83.0	24.29	28.95	83.9	1	0	0
United States	100.00	100.0	100.00	100.00	100.0	1	1	0

Source: Frederick T. Schut and Peter A. G. VanBergeijk, “International Price Discrimination: The Pharmaceutical Industry,” *World Development*, Vol. 14, No. 9, p. 1144.
Datafile = DRUGS5

5.8 Appendix: Econometric Lab #3

This lab focuses on hypothesis testing. You will estimate models of life expectancy at birth across the 50 states and the District of Columbia using economic and demographic variables. The data are in the dataset LIFE5 on the textbook’s website and include the following variables:

Table 5.3 Variable Listing

Variable	Description
lifeexpect_i	Life expectancy at birth, in years, in state i, 2010
medinc_i	The median household income in state i (thousands of dollars), 2010
uninsured_i	The percentage of the population (aged 0–64) in state i that was without health insurance coverage, 2008–2010
smoke_i	The percentage of adults in state i who smoked, 2006–2012
obesity_i	The percentage of adults in state i who were obese (Body Mass Index greater than or equal to 30), 2006–2012
teenbirth_i	The number of births to teenaged mothers in state i per 1,000 females aged 15 to 19 years, 2010
gunlaw_i	A dummy variable = 1 if state i had a firearm law protecting children, 0 otherwise, 2010
metro_i	The percentage of the population in state i that lived in a metropolitan statistical area, 2010

Step 1: Specify the Model

Specify (i.e., write out) a linear regression equation for **lifeexpect** with all seven independent variables included, using the format of Equation 5.1 in the text. Use proper subscripts and Greek letters where appropriate.

Step 2: Hypothesize the Signs of the Coefficients

For all seven independent variables, hypothesize the sign of each regression coefficient.

Step 3: Summary Statistics

Check the means, maximums, and minimums for each of the variables. Do you spot any obvious anomalies? If so, what are they? If you see no anomalies, go on to Step 4.

Step 4: Estimation

Run the regression using all seven independent variables and print out your regression results.

Step 5: Hypothesis Testing (*t*-statistics)

Test the slope coefficients of **smoke**, **teenbirth**, **medinc**, and **uninsured** at the 5-percent level of significance using the *t*-table in the textbook. Show your null and alternative hypotheses and list the critical *t*-statistic used for each hypothesis test. For which coefficients can you reject the null hypothesis?

Step 6: Hypothesis Testing (*p*-values)

Test the slope coefficients of **gunlaw**, **metro**, and **obesity** at the 5-percent level of significance using *p*-values. List the *p*-value used for each test. For which coefficients can you reject the null hypothesis?

Step 7: Overall *F*-test

Use the overall *F*-statistic to test whether the regression is significant at the 5-percent level. Show your null and alternative hypotheses and your decision rule, and use the *F*-table.

Step 8: Drawing Conclusions

The absolute value of the coefficient of **gunlaw** is much larger than the absolute value of the coefficient of **smoke**. Does this mean that passing a gun law to protect children will have a bigger impact on life expectancy than reducing smoking by three percentage points? Explain.

Chapter 6

Specification: Choosing the Independent Variables

Before any equation can be estimated, it must be specified. *Specifying* an econometric equation consists of three parts: choosing the correct independent variables, the correct functional form, and the correct form of the stochastic error term.

A **specification error** results when any one of these choices is made incorrectly. This chapter is concerned with only the first of these, choosing the variables; the second and third choices will be taken up in later chapters.

The fact that researchers can decide which independent variables to include in regression equations is a source of both strength and weakness in econometrics. The strength is that the equations can be formulated to fit individual needs, but the weakness is that researchers can estimate many different specifications until they find the one that "proves" their point, even if many other results disprove it. A major goal of this chapter is to help you understand how to choose variables for your regressions without falling prey to the various errors that result from misusing the ability to choose.

The primary consideration in deciding whether an independent variable belongs in an equation is whether the variable is essential to the regression on the basis of theory. If the answer is an unambiguous yes, then the variable definitely should be included in the equation, even if it seems to be lacking in statistical significance. If theory is ambivalent or less emphatic, a dilemma arises. Leaving a relevant variable out of an equation is likely to bias the remaining estimates, but including an irrelevant variable leads to higher variances of the estimated coefficients. Although we'll develop statistical tools to help us deal with this decision, it's difficult in practice to be sure that a variable is relevant, and so the problem often remains unresolved.

We devote the fourth section of the chapter to specification searches and the pros and cons of various approaches to such searches. For example, poorly done specification searches often cause bias or make the usual tests of significance inapplicable. Instead, we suggest trying to minimize the number of regressions estimated and relying as much as possible on theory rather than statistical fit when choosing variables. There are no pat answers, however, and so the final decisions must be left to each individual researcher.

6.1 Omitted Variables

Suppose that you forget to include one of the relevant independent variables when you first specify an equation (after all, no one's perfect!). Or suppose that you can't get data for one of the variables that you *do* think of. The result in both these situations is an **omitted variable**, defined as an important explanatory variable that has been left out of a regression equation.

Whenever you have an omitted (or *left-out*) variable, the interpretation and use of your estimated equation become suspect. Leaving out a relevant variable, like price from a demand equation, not only prevents you from getting an estimate of the coefficient of price but also usually causes bias in the estimated coefficients of the variables that are in the equation.

The bias caused by leaving a variable out of an equation is called **omitted variable bias**. In an equation with more than one independent variable, the coefficient β_k represents the change in the dependent variable Y caused by a one-unit increase in the independent variable X_k, holding constant the other independent variables in the equation. If a variable is omitted, then it is not included as an independent variable, and it is not held constant for the calculation and interpretation of $\hat{\beta}_k$. This omission can cause bias: It can force the expected value of the estimated coefficient away from the true value of the population coefficient.

Thus, omitting a relevant variable is usually evidence that the entire estimated equation is suspect, because of the likely bias in the coefficients of the variables that remain in the equation. Let's look at this issue in more detail.

The Consequences of an Omitted Variable

What happens if you omit an important variable from your equation (perhaps because you can't get the data for the variable or didn't even think of the variable in the first place)? The major consequence of omitting a relevant independent variable from an equation is to cause bias in the regression coefficients that remain in the equation. Suppose that the true regression model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{6.1}$$

where ϵ_i is a classical error term. If you omit X_2 from the equation, then the equation becomes:

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \epsilon_i^* \tag{6.2}$$

where ϵ_i^* equals:

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{2i} \tag{6.3}$$

because the stochastic error term includes the effects of any omitted variables, as mentioned in Section 1.2. Why does Equation 6.2 include β_0^* and β_1^* instead of β_0 and β_1? The answer lies in the meaning of a regression coefficient. β_1 is the impact of a one-unit increase in X_1 on Y, *holding X_2 constant*, but X_2 isn't in Equation 6.2, so OLS can't hold it constant. As a result, β_1^* is the impact of a one-unit increase in X_1 on Y, *not holding X_2 constant*.

From Equations 6.2 and 6.3, it might seem as though we could get unbiased estimates even if we left X_2 out of the equation. Unfortunately, this is not the case,[1] because the included coefficients almost surely pick up some of the effect of the omitted variable and therefore will change, causing bias. To see why, take another look at Equations 6.2 and 6.3. Most pairs of variables are correlated to some degree, so X_1 and X_2 almost surely are correlated. When X_2 is omitted from the equation, the impact of X_2 goes into ϵ^*, so ϵ^* and X_2 are correlated. Thus if X_2 is omitted from the equation and X_1 and X_2 are correlated, both X_1 and ϵ^* will change when X_2 changes, and the error term will no longer be independent of the explanatory variable. That violates Classical Assumption III!

In other words, if we leave an important variable out of an equation, we violate Classical Assumption III (that the explanatory variables are independent of the error term), unless the omitted variable is uncorrelated with all the included independent variables (which is extremely unlikely). In general,

[1] To avoid bias, X_1 and X_2 must be perfectly uncorrelated in the sample—an extremely unlikely result.

when there is a violation of one of the Classical Assumptions, the Gauss–Markov Theorem does not hold, and the OLS estimates are not BLUE. Given linear estimators, this means that the estimated coefficients are no longer unbiased or are no longer minimum variance (for all linear unbiased estimators), or both. In such a circumstance, econometricians first determine the exact property (unbiasedness or minimum variance) that no longer holds and then suggest an alternative estimation technique that might be better than OLS.

An omitted variable causes Classical Assumption III to be violated in a way that causes bias. Estimating Equation 6.2 when Equation 6.1 is the truth will cause bias. This means that:

$$E(\hat{\beta}_1^*) \neq \beta_1 \tag{6.4}$$

Instead of having an expected value equal to the true β_1, the estimate will compensate for the fact that X_2 is missing from the equation. If X_1 and X_2 are correlated and X_2 is omitted from the equation, then the OLS estimation procedure will attribute to X_1 variations in Y that are actually caused by X_2, and a biased estimate of β_1 will result.

To see how an omitted variable can cause bias, let's look at an extremely early application of regression analysis.[2] During World War II, the Allies were interested in improving the accuracy of their bombers, so they estimated an equation where the dependent variable was bomber accuracy and the independent variables included such things as the speed and altitude of the bombing group and the amount of enemy fighter opposition. As expected, the estimated coefficients supported the hypotheses that higher speeds and higher altitudes led to larger aiming errors, but the researchers were shocked to discover that more enemy fighter opposition appeared to improve the accuracy of the pilot and bombardier! What was going on?

The answer is omitted variable bias. It turns out that the equation didn't include a variable for cloud cover over the target, and cloud cover typically prevented enemy fighters from flying. When it was cloudy, the bombers couldn't see the ground and made large errors, but OLS attributed these errors to the lack of enemy fighter opposition because there was no variable for cloud cover in the equation and because few fighters could fly when it was cloudy. Put differently, the coefficient of enemy fighters picked up the impact of the omitted variable of cloud cover because the two variables were highly correlated. This is omitted variable bias!

[2] Adapted from Frederick Mosteller and John Tukey, *Data Analysis and Regression: A Second Course in Statistics* (Reading, MA: Addison-Wesley, 1977), p. 318.