

Multivariate Regression and the Omitted Variable Bias

“All Else Equal”

- Each of the predictions of theory include the phrase “all else equal.”
- Sometimes the same phrase is written “holding all other factors constant” or “ceteris paribus” (from Latin).
- In any case, the phrase emphasizes the point that there are always multiple independent variables that affect the same dependent variable.
- Because the predictions of theory assume that the other factors are held constant, a test of the theory should account for those factors as well.

Multivariate Regression Analysis

- The univariate regression analysis implicitly assumed that no other factors influenced the dependent variable.

For example:

Dependent variable = Quit rates in firms;
Independent Variable = Average Wage

- To account for other factors, one can simply add an additional independent variable into the regression equation.

For example: The average age of the workers in a firm might affect the quit rate; i.e. Older workers are less likely to quit their jobs than young workers.

Data

- Firms 1-3 have employees whose average age is less than 40 while firms 4-6 have employees whose average age is greater than 40.

Data

- Firms 1-3 have employees whose average age is less than 40 while firms 4-6 have employees whose average age is greater than 40.

Firm	Average Wage	Quit Rate	Average Age	A_i
1	4	40%	< 40	0
2	6	35	< 40	0
3	8	30	< 40	0
4	8	20	> 40	1
5	10	15	> 40	1
6	12	10	> 40	1

Note: Dummy Variables

- A_i is called a dummy variable because it only takes on the values of 0 and 1.
- In this case, A_i takes on a value of 1 when average age is greater than 40 and 0 when average age is less than 40.
- Dummy variables can be thought of as true/false indicators where 1 is true and 0 is false.
- A_i might be referred to as the “older workers dummy” because it takes on a value of 1 when “older workers” is true.

Multivariate Regression Analysis continued...

- Univariate case:

$$Q_i = \alpha_0 + \alpha_1 W_i + \epsilon_i$$

- Adding the older workers dummy variable gives a new equation describing the data:

$$Q_i = \alpha'_0 + \alpha'_1 W_i + \alpha'_2 A_i + \epsilon'_i$$

Note 1: The coefficient on wage and the constant (α'_0) both include the symbol ' ; because they are theoretically different from the univariate regression:

$$\alpha'_0 \neq \alpha_0 \text{ and } \alpha'_1 \neq \alpha_1 \text{ and } \epsilon'_i \neq \epsilon_i$$

Note 2: The dummy variable A_i takes advantage of the identity and zero properties of multiplication: $X*0 = 0$ and $X*1 = X$.

When $A_i = 1$ (age > 40), $\alpha'2 A_i = \alpha'2$. When $A_i = 0$ (age < 40), $\alpha'2 A_i = 0$. Therefore, the term $\alpha'2 A_i$ simply adds $\alpha'2$ to the quit rate when age > 40.

Omitted Variable Bias

- Using ordinary least squares, we estimate the equation:

$$E[Q_i | W_i, A_i] = 60 - 4 W_i$$

- The estimates, $\hat{\alpha}_0$, $\hat{\alpha}_1$ are 60, -4, respectively.

- Also, using ordinary least squares, we estimate the equation:

$$E[Q_i | W_i, A_i] = 50 - 2.5 W_i - 10 A_i$$

- The estimates, $\hat{\alpha}'_0$, $\hat{\alpha}'_1$ and $\hat{\alpha}'_2$ are 50, -2.5 and -10, respectively.

- The estimate $\hat{\alpha}'_2 = -10$ says that, all else equal, a firm with average age greater than 40 will have 10% fewer quits per year than a firm with average age less than 40.

- Let's assume that the estimates are statistically significant.

- The univariate regression implicitly and incorrectly assumed that the coefficient on the older worker dummy was equal to zero.

- Because of this incorrect assumption the estimate of the effect of wage on the quit rate was biased. We call this an **omitted variable bias**.

- The biased regression (sometimes called the short regression) gave an estimate of $\hat{\alpha}_1 = -4$, while the unbiased (long) regression gave an estimate of $\hat{\alpha}'_1 = -2.5$.

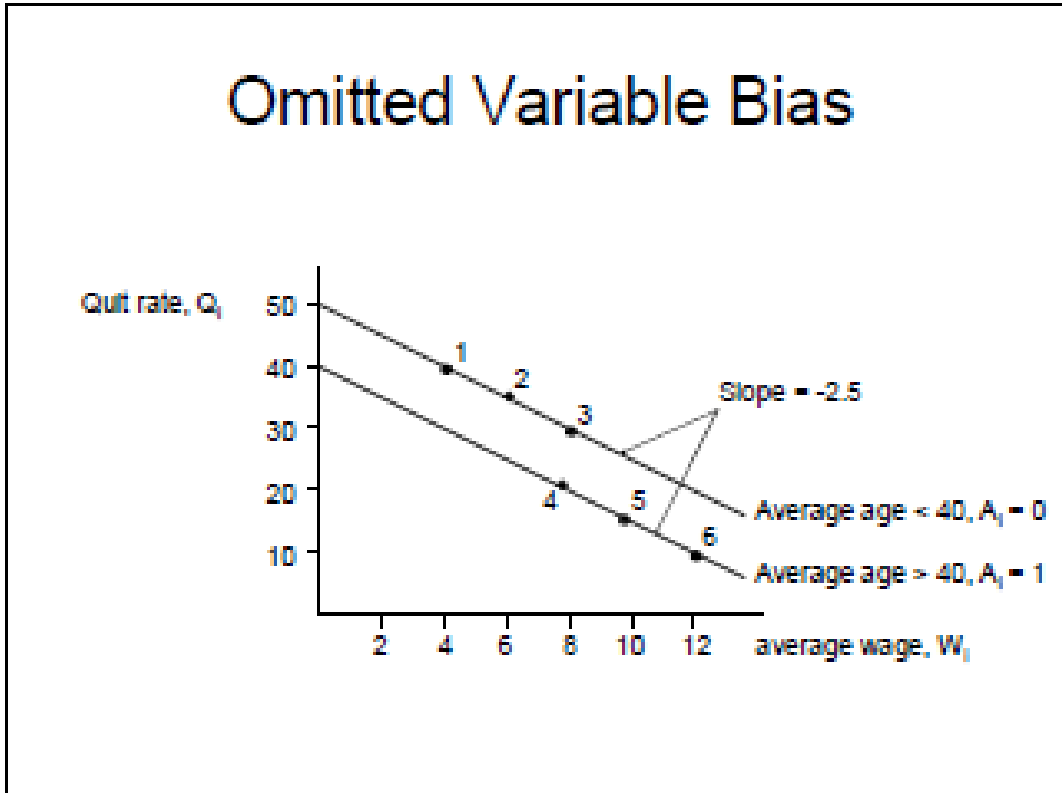
- In this case, the omitted variable caused the coefficient on wage to be overestimated, because $|\hat{\alpha}_1| > |\hat{\alpha}'_1|$. (The effect of a wage change was overestimated.)

- An overestimation is often called an upward bias. The estimate is biased away from zero.

- An underestimation may be called a downward bias, or an attenuation bias, meaning that the estimate is biased toward zero.

- We can graph (see below) the multivariate regression by drawing one line for firms where $A_i = 1$ and a separate line for firms where $A_i = 0$ different line for firms.

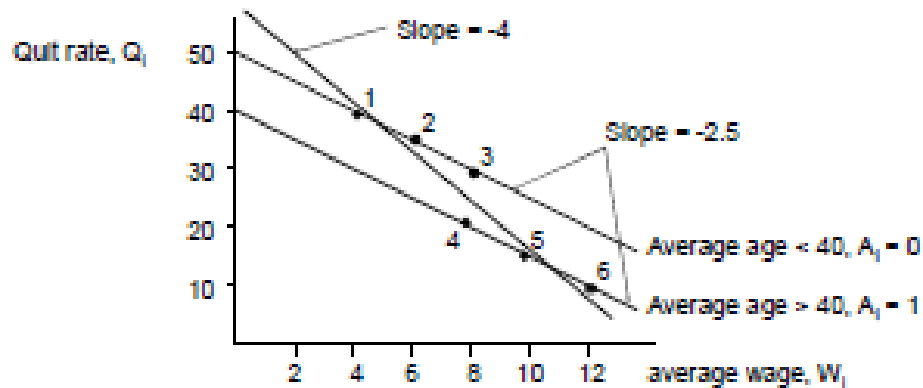
- Notice that at any wage, the effect of going from average age < 40 to average age > 40 is a drop in the quit rate of 10%.



- Notice also that the effect of increasing the average wage by \$1, (a drop in the quit rate of 2.5%) is the same for both older worker firms and younger worker firms. This is by assumption.

Omitted Variable Bias

- When the univariate best fit line is added, it is clear, visually, that the multivariate estimate fits the data better.



Omitted Variable Bias continued...

- Omitted variable bias occurs when

1) the omitted variable has an effect on the dependent variable,

AND

2) the omitted variable is correlated with the independent variable of interest.

- In our case, both were true.

1) The omitted variable, average age, caused the dependent variable, quit rates, to drop by 10% when average age was greater than 40.

2) The omitted variable, average age, tends to be high when the independent variable of interest, average wage, is high, so they are correlated.

- How do you know if the estimate is biased upward or downward?

Upward or Downward Bias

- Three variables: the dependent variable, the independent variable of interest, and the omitted variable.

- Three pairs of variables:

dependent variable \Leftrightarrow independent variable of interest

dependent variable \Leftrightarrow omitted variable

omitted variable \Leftrightarrow independent variable of interest

- The relationship between the variables in each pair, tells you whether the estimate will be upward or downward biased.

dependent variable \Leftrightarrow independent variable of interest

dependent variable \Leftrightarrow omitted variable

omitted variable \Leftrightarrow independent variable of interest

(i) + + + If there is a positive correlation within each pair, the estimate will be upward biased.

(ii) + + – If there is a positive correlation within two of the pairs and a negative correlation within one pair, the estimate will be downward biased.

(iii) + – – If there is a positive correlation within one of the pairs and a negative correlation within two pairs, the estimate will be upward biased.

(iv) – – – If there is a negative correlation within each pair, the estimate will be downward biased.

In the example:

dependent variable \Leftrightarrow independent variable of interest

(Quit Rate) – negative correlation (Average Wage)

dependent variable \Leftrightarrow omitted variable

(Quit Rate) – negative correlation (Average Age)

omitted variable \Leftrightarrow independent variable of interest

(Average Age) + positive correlation (Average Wage)

- One positive correlation and two negative correlations imply that estimate of the coefficient on Wage in the short regression was upward biased (as we saw previously).