Ch. 6 Specification: Independent Variables

- **Omitted Variables**
  - important (relevant) explanatory variables left out of an estimated regression function.
  1) consequence: may cause bias in the remaining estimated regression coefficients (called **Omitted variable bias**).
  2) An example on production function:

\[
\text{Output} = \beta_0 + \beta_1 \text{Labor} + \beta_2 \text{Capital} + \epsilon.
\]

If Capital is left out, the OLS would attribute to labor the increase in output actually caused by capital as labor and capital are correlated.

3) \( E(\hat{\beta}_1) = \beta_1 + \beta_2 \cdot \alpha_1 \) where \( \alpha_1 \) is in the regression of \( x_2 \) on \( x_1 \), i.e., \( x_2 = \alpha_0 + \alpha_1 x_1 + u \). The bias is \( \beta_2 \cdot \alpha_1 \) for the estimate \( \hat{\beta}_1 \).

4) An numerical example: The annual consumption of chicken

\[
\hat{y}_t = 31.5 - 0.73PC_t + 0.11PB_t + 0.23YD_t
\]

\[
(0.08) \quad (0.05) \quad (0.02)
\]

\[
t = -9.12 \quad 2.50 \quad 14.22
\]

\[
\bar{R}^2 = 0.986 \quad n = 44,
\]

where \( Y_t \) = per capita chicken consumption (in pounds) in year \( t \)

\( PC_t \) = the price of chicken (in cents per pound) in year \( t \)

\( PB_t \) = the price of beef (in cents per pound) in year \( t \)

\( YD_t \) = U.S. per capita disposable income (in hundreds of dollars) in year \( t \).

a) This is a demand for chicken equation that includes the prices of chicken and a close substitute (beef) and an income variable.

b) If the equation is estimated without the price of the substitute, we obtain

\[
\hat{y}_t = 32.9 - 0.70PC_t + 0.27YD_t
\]

\[
(0.08) \quad (0.01)
\]

\[
t = -8.33 \quad 45.91
\]

\[
\bar{R}^2 = 0.984 \quad n = 44.
\]

c) By comparing Eqs.(1) and (2), one can see the impact of dropping the important beef price variables on the estimated equations.

- **Including Irrelevant Variables**
  - the inclusion of explanatory variables that do not belong to the regression equation.
1) Consequences: do not cause bias, but will increase the variances of the estimated coefficients of the included variables.

2) A numerical example: $y_t$ the annual consumption of chicken; the original estimated equation is Eq.(1) above. Suppose now we include the interest rate (irrelevant) variable $R_t$,

$$\hat{y}_t = 30.0 - 0.73PC_t + 0.12PB_t + 0.22YD_t + 0.17R_t$$

$$\begin{aligned}
(0.08) & \quad (0.06) & \quad (0.02) & \quad (0.21) \\
\hat{t} &= -9.10 \quad 2.08 \quad 11.05 \quad 0.82
\end{aligned}$$

$$\hat{R}^2 = 0.985 \quad n = 44,$$

a) $\hat{R}^2$ has fallen slightly.
b) Slight increases in the standard errors.
c) The $t$-value for the irrelevant variable $R_t$ is very small.

- The following example demonstrates that dropping variables with small $t$-values might not be a good practice. Sound economic theory is important!

1) The demand for Brazilian coffee in US,

$$\hat{COFFEE} = 9.1 + 7.8P_{bc} + 2.4P_t + 0.0035Y_d$$

$$\begin{aligned}
(15.6) & \quad (1.2) & \quad (0.001) \\
\hat{t} &= 0.5 \quad 2.0 \quad 3.5
\end{aligned}$$

$$\hat{R}^2 = 0.60 \quad n = 25,$$

where $COFFEE =$ the demand for Brazilian coffee in US

- $P_{bc}$ = the real price of Brazilian coffee
- $P_t$ = the real price of tea
- $Y_d$ = real disposable income in US.

--- $P_{bc}$ appears to have an insignificant coefficient with an unexpected sign. One may believe from this that Brazilian coffee is perfectly inelastic, i.e., zero coefficient.

2) Decide to reestimate the same equation without the price variable:

$$\hat{COFFEE} = 9.3 + 2.6P_t + 0.0036Y_d$$

$$\begin{aligned}
(1.0) & \quad (0.0009) \\
\hat{t} &= 2.6 \quad 4.0
\end{aligned}$$

$$\hat{R}^2 = 0.61 \quad n = 25.$$ 

a) $\hat{R}^2$ increases slightly when $P_{bc}$ is dropped.
b) $SE$'s are slightly decreased.
c) The estimated coefficients change only a small amount.

--- one might attempt to conclude that the demand for Brazilian coffee is indeed perfectly price inelastic.

The price of Brazilian coffee is irrelevant.
3) A better theoretical model: Include competition from other kinds of coffee

\[ \hat{\text{COFFEE}} = 10.0 + 8.0 P_{cc} - 5.6 P_{bc} + 2.6 P_{t} + 0.0030 Y_d \]

\[ (4.0) \quad (2.0) \quad (1.3) \quad (0.0010) \]

\[ t = 2.0 \quad -2.8 \quad 2.0 \quad 3.0 \]

\[ \bar{R}^2 = 0.65 \quad n = 25, \]

where \( P_{cc} \) is the price of Colombian coffee.

a) Both the \( t \)-values of the coffees are significant and have the theoretical expected signs. Colombian coffee is a substitute for Brazilian coffee.

b) \( \bar{R}^2 \) is better.

\[ \text{--- Conclusion: theoretical considerations should never be discarded.} \]

- An Example: Choosing Independent Variables

1) Try to explain GPA of students. Data are obtained on the following variables (25 observations)

- \( GPA \): the cumulative college grade point average
- \( HGPA \): the cumulative high school grade point average
- \( MSAT \): the highest score earned on the math section of the SAT
- \( VSAT \): the highest score earned on the verbal section of the SAT
- \( SAT = MSAT + VSAT \)
- \( GREK \): a dummy variable, 1 if a fraternity member; 0, otherwise.
- \( HRS \): the average number of hours spent studying per course per week
- \( PRIV \): a dummy variable, 1 if graduated from a private high school; 0, otherwise.
- \( JOCK \): a dummy variable, 1 if a member of an athletic team for at least one season; 0 otherwise.
- \( InEX \): the natural log of the number of full courses completed in college

Assuming \( GPA \) is the dependent variable, which independent variables would be used?

2) (production function) specification:

\[ \text{grades} = f(\text{ability}, \text{hardwork}, \#\text{courses} \text{taken}) + \epsilon \]

So, choose the explanatory variables: \( HGPA, HRS, InEX \).

3) How are the other variables?

a) SAT’s: Once we know HGPA, SAT’s could be redundant.

b) \( JOCK \) and \( GREK \): Once we know how many hours student spends studying, don’t care what that student does with the rest of his/her time.

c) \( PRIV \) may only have a minor effect.

4) With 25 students, we obtain:
\[ \hat{GP}_A = -0.26 + 0.49 HGPA + 0.06 HRS + 0.42 \ln EX \]

\[
\begin{array}{ccc}
(0.21) & (0.02) & (0.14) \\
2.33 & 3.00 & 3.00 \\
\end{array}
\]

\[ R^2 = 0.585 \quad F = 12.3 \quad n = 25. \]

5) If SAT is included,

\[ \hat{GP}_A = -0.92 + 0.47 HGPA + 0.05 HRS + 0.44 \ln EX + 0.00060 SAT \]

\[
\begin{array}{cccc}
(0.22) & (0.02) & (0.14) & (0.00064) \\
2.12 & 2.50 & 3.12 & 0.93 \\
\end{array}
\]

\[ R^2 = 0.583 \quad F = 9.4 \quad n = 25. \]