Introductory Econometrics

Introduction to Econometrics

Summation

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

The symbol means "the sum of" and the letter i is the index of summation (arbitrary letter that may appear as t, j or k).

The numbers 1 and n are the lower and upper limits of the summation.

Properties of the Summation Operator

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_3$$

$$\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i \quad \longrightarrow \quad a \text{ is a constant}$$

$$\sum_{i=1}^{n} a = a + a + \cdots a = na$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} (a x_i + b y_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

Use of Multiple Signs in One Expression:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i, y_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_i + y_j)$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i, y_j) = \sum_{i=1}^{m} [f(x_i, y_1) + f(x_i, y_2) + \cdots + f(x_i, y_n)]$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i, y_j) = \sum_{j=1}^{n} \sum_{i=1}^{m} f(x_i, y_j)$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_i y_j = \left(\sum_{i=1}^{m} x_i \right) \left(\sum_{j=1}^{n} y_j \right)$$

Definitions

Integers: Are the whole numbers (without a fractional component). Eg: ∓ 1, 2, 3

Rational Numbers: Can be written as the quotient of 2 integers, *a/b* where $b \neq 0$

Real Number: Is a quantity that can be represented by a point on a continuous line (whole, rational, irrational).

Real numbers that cannot be expressed as a quotient of 2 integers are called **irrational** numbers. Eg: $\sqrt{2}\ or\ e$

Numbers like $\sqrt{-2}$ are not real numbers and they are called **imaginary numbers**.

The **absolute value** of a number is denoted $|a|$

Inequalities

Inequalities among numbers obey certain rules.

If a < b, then a + c < b + c

If a < b, then $\begin{cases} ac < bc & \text{if } c > 0 \\ ac > bc & \text{if } c < 0 \end{cases}$

If a < b and b < c, then a > c

Exponents

Exponents are defined as follows:

$$x^n = x \times x \times \ldots \times x \ (n \text{ terms}) \text{ if } n \text{ is a positive integer}$$

- $x^0 = 1$ if $x \neq 0$
- 0^0 is "undefined"

Some common rules:

$x^{-n} = \frac{1}{x^n}\ if\ x \neq 0$, $x^{1/n} = \sqrt[n]{x}$, $x^{m/n} = \left(x^{1/n}\right)^m$

$x^a x^b = x^{a+b}$, $\frac{x^a}{x^b} = x^{a-b}$, $\left(\frac{x}{y}\right)^a = \frac{x^a}{y^a}$

$(xy)^a = x^a \times y^a$

Scientific Notation

A number in scientific notation is written as a number between 1 and 10 and multiplied by a power of 10 as in:

$$5.1 \times 10^5 = 510{,}000 \quad \text{and} \quad 0.00000034 = 3.4 \times 10^{-7}$$

Example:

$$\begin{aligned} 510{,}000 \times 0.00000034 &= (5.1 \times 10^5) \times (3.4 \times 10^{-7}) \\ &= (5.1 \times 3.4) \times (10^5 \times 10^{-7}) \\ &= 17.34 \times 10^{-2} \\ &= 0.1734 \end{aligned}$$

$$\frac{510000}{0.00000034} = \frac{5.1 \times 10^5}{3.4 \times 10^{-7}} = 1.5 \times 10^{12}$$

- In computer outputs, exponents are often written as follows:
  $5.1 \times 10^5 = 5.1E5$
  $3.4 \times 10^{-7} = 3.4E-7$

Logarithm

The logarithm of a number, say A, with base B, is the number of times B has to be multiplied (ie the exponent or power) to get A.

$\log_{10}(100) = 2$ because $10^2 = 100$.
$\log_e(2.7182318\ldots) = 1$ because $e = 2.7182318\ldots$
$\log_e(e^b) = b$

Logarithms with base e are called **natural logarithms** and often written as *ln.* We will only use natural logarithms in this unit.

$B^c = A \iff c = \log_B(A)$

$B = e \qquad c = \log_e(A) = ln(A)$

$\Updownarrow$

$A = e^c = exp(c)$

- $ln(e^b) = b$
- $ln(xy) = ln(x) + ln(y)$
- $ln\left(\frac{x}{y}\right) = ln(x) - ln(y)$
- $\ln(x^a) = a ln(x)$

By definition:

$$x = e^{ln(x)} = \exp[ln(x)]$$

When there is an **exponential function** with a complicated exponent, the notation **exp** is often used so that e(.) = exp (.)

The above shows that the exponential function is the **antilogarithm**. That is the exponential function (antilog) of *ln(x)* is *x*.

Proportional (or Relative) Change

A proportional (or relative) change in y is defined as the change in y as a proportion of the value of y before the change.

$$\text{Relative change in } y = \frac{y_1 - y_0}{y_0} = \frac{\Delta y}{y_0}$$

A proportional change may be expressed as a percentage **(percentage change):**

$$\text{Percentage change } y = 100 \times \frac{y_1 - y_0}{y_0} = \%\Delta y$$

Note that a change in the natural log of y is approximately a proportional change in y, provided that the change is small relative to the initial value. That is:

$$\Delta ln(y) = ln(y_1) - \ln(y_0) \approx \Delta y / y_0 = (y_1 - y_0) / y_0$$

For example if $y_0 = 1$

y_1	$\%\Delta y$	$100\Delta \ln(y)(\%)$	Approximation error (%)
1.01	1.00	0.995	0.50
1.05	5.00	4.88	2.48
1.10	10.00	9.53	4.92
1.15	15.00	13.98	7.33
1.20	20.00	18.23	9.70
1.25	25.00	22.31	12.04

Linear Relationships

The standard form for a linear relationship is:

$$y = mx + b \qquad = f(x)$$

The slope of m is the **marginal effect** of a change in x on y:

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} \qquad \Delta y = m\Delta x$$

The intercept parameter, b, indicates where the line crosses the vertical axis- that is, it is the value of y when x is zero.

The first derivative of f(x) measures the slope of the tangent to f(x):

$$\frac{dy}{dx} = m$$

Δy and Δx represent a discrete change in y and x, respectively. ie:

$\Delta y = y_1 - y_2$ and $\Delta x = x_1 - x_2$

while dy and dx represent infinitesimal changes in y and x respectively.

When the function is linear:

$$\frac{dy}{dx} = \frac{\Delta y}{\Delta x}$$

Elasticity:

An elasticity is the percentage change in one variable associated with 1% increase in another variable.

For a linear relationship, the elasticity of y with respect to a change in x is:

$$\varepsilon_{yx} = \frac{\%\Delta y}{\%\Delta x} = \frac{100(\Delta y/y)}{100(\Delta x/x)} = \frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x} \times \frac{x}{y} = slope \times \frac{x}{y}$$

Example

If $y = 5 + 2x$, elasticity at point $x = 2$, $y = 9$

- Slope × $(x/y) = 2 \times (2/9) = 4/9 = 0.44$ (2dp)
- That is at that point ($x = 2$, $y = 9$) a 1% increase in x is associated with 0.44% change in y.
- At $x = 2$ a 1% increase amounts to a change by 0.02 ($\Delta x = 0.01 \times 2$).
- If x increases to 2.02, the value of y increases to $(9 + 0.04 = 9.04)$. The relative change in y is $\Delta y/y = 0.04/9 = 0.0044$.
- The percentage change in y is

$$\%\Delta y = 100(\Delta y/y) = 100 \times 0.0044 = 0.44\%$$

Non-Linear Relationship

The slope (derivative) of a nonlinear function is not constant, but changes as x changes and must be determined at each point.

The slope of a curve at a point is the slope of the line that is tangent to the curve at that point.

Quadratic Function:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2$$

If $\beta_3 > 0$: the curve (parabola) is U-shaped and it has a minimum.

If $\beta_3 < 0$: the curve (parabola) has a maximum.

The maximum or minimum occurs at the point where the following holds:

$$\frac{dy}{dx} = \beta_2 + 2\beta_3 x = 0 \ or \ \ x = -\beta_2/(2\beta_3)$$

Cubic Function:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3$$

Cubic functions have one **inflection point** where the function crosses its tangent line and changes from concave to convex or vice versa.

Cubic functions can be used for the total cost or total product function in economics. Then, the first derivative of the total cost is marginal cost and the first derivative of the total product is marginal product.

If the “total” curves are cubic, then the “marginal” curves are quadratic (u-shaped curve for marginal cost and inverted u-shape for marginal product).

Log Function:

$$ln(\text{salary}) = \beta_1 + \beta_2 \ln(sales)$$

The elasticity of salary with respect to sales is β_2 and it is fixed for all values of sales. The logarithmic changes are always percentage changes.

$$\beta_2 = \frac{\Delta ln(salary)}{\Delta ln(sales)} = \frac{100 \times \frac{\Delta salary}{salary}}{100 \times \frac{\Delta sales}{sales}}$$

Percentage change of salary

Percentage change of sales

On the other hand, the effect of a marginal change in sales on salary is not constant but changes with sales:

$$\frac{d(Salary)}{d(Sales)} = \beta_2 \frac{Salary}{Sales}$$

- Example: $ln(salary) = 6.52 + 0.324 \ln(sales)$

  1% sales increase results in 0.324% salary increases

- The log-log form postulates a constant elasticity model, whereas the semi-log form assumes a semi-elasticity model.

Log-Linear Function:

- Example: $\ln(wage) = \beta_1 + \beta_2 educ$

relative change of wage

$$\beta_2 = \frac{\Delta \ln(wage)}{\Delta educ} = \frac{1}{wage}\frac{\Delta wage}{\Delta educ} = \frac{\frac{\Delta wage}{wage}}{\Delta educ}$$

if years of education are increased by one year

- Elasticity of *wage* w.r.t. *educ* = $\beta_2 \times$ educ

$$Ln(wage) = 0.767 + 0.056 educ$$

Wage increases by 5.6% for every additional year of education.

$$Food_{Exp} = \beta_1 + \beta_2 \, ln(Income)$$

The slope of this function is $\frac{dy}{dx} = \beta_2 / income$ and it changes with income

- Interpretation of β_2: a 1% increase in income leads to a $\frac{\beta_2}{100}$ unit change in food expenditure.

$$Food_{Exp} = -97.19 + 137.17 \, ln(Income)$$

- 1% increase in income will increase food expenditure by approximately \$1.37 per week, or a 10% increase in income will increase food expenditure by approximately \$13.72.

Some useful Functions, their Derivatives, Elasticities and other interpretation

Interpretation

Name	Function	Slope = dy/dx	Elasticity
Linear	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Quadratic	$y = \beta_1 + \beta_2 x^2$	$2\beta_2 x$	$(2\beta_2 x)\frac{x}{y}$
Cubic	$y = \beta_1 + \beta_2 x^3$	$3\beta_2 x^2$	$(3\beta_2 x^2)\frac{x}{y}$
Log-Log	$\ln(y) = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{y}{x}$	β_2
Log-Linear	$\ln(y) = \beta_1 + \beta_2 x$	$\beta_2 y$	$\beta_2 x$
	or, a 1 unit change in x leads to (approximately) a 100 β_2% change in y		
Linear-Log	$y = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$
	or, a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y		

The Econometric Model

In economics, ideas of the relationships between economic variables are expressed using a mathematical function. These functions have unknown **parameters** and economics provides systematic methods to estimate those parameters and answer questions like:

- Can you predict the sales of the product?
- Can you estimate the effect on our sales if our competitor lowers their price by $1?
- Can you test whether our new advertising campaign is actually increasing our sales?

An econometric model consists of a systematic (or theoretical) part and a random and unpredictable component e called a **random error.**

$$Q^d = f\left(P, P^s, P^c, INC\right) + e$$
$$f\left(P, P^s, P^c, INC\right) = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 INC$$
$$Q^d = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 INC + e$$

These functional form represents the relationship between the variables that we assume is true (or close to the true relationship).

The coefficients β_1, β_2 are unknown parameters of the model that we estimate using economic data and an econometric technique.

The systematic portion is the part we obtain from economic theory and includes an assumption about the function form.

The random component represents a "noise" component which we represent using the random variable e.

We use the econometric model as a basis for statistical inference. The ways with which statistical inferences are drawn include:

1. Estimating economic parameters, such as elasticities, using econometric methods
2. Predicting economic outcomes such as enrolment in 2 year colleges in US for next 10 years
3. Testing economic hypotheses such as the question of whether newspaper advertising is better than store displays for increasing sales

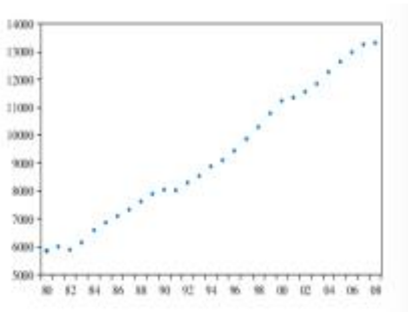
Economic Data Types

Economic data comes in a variety of flavours:

- Data may be collected at various levels of aggregation: Micro or Macro
- Data may also represent a flow or a stock:
    - Flow: measured over a period of time
    - Stock: measured at a particular point in time
- Data may be qualitative or quantitative:
    - Quantitative: Expressed as numbers
    - Qualitative: Expressed as an “either or” situation

A **time series** is data collected over discrete intervals of time. The key feature is that the same economic quantity is recorded at regular time intervals.

Year	GDP
2001	11347.2
2002	11553.0
2003	11840.7
2004	12263.8
2005	12638.4
2006	12976.2
2007	13254.1
2008	13312.2



A **cross-section of data** is collected across sample units in a particular time period. The sample units are individual entities and may be firms, persons, states, households, countries.

A **panel of data**, also known as “longitudinal” data has observations on individual micro-units who are followed over time.

- The key aspect of panel data is that we observe each micro-unit for a number of time periods
- If we have the same number of time period observations for each micro-unit, we have a **balanced panel**
- Usually the number of time series observations is small relative to the number of micro-units but not always

FIRM	*YEAR*	*PROD*	*AREA*	*LABOR*	*FERT*
1	1990	7.87	2.50	160	207.5
1	1991	7.18	2.50	138	295.5
1	1992	8.92	2.50	140	362.5
1	1993	7.31	2.50	127	338.0
1	1994	7.54	2.50	145	337.5
1	1995	4.51	2.50	123	207.2
1	1996	4.37	2.25	123	345.0
1	1997	7.27	2.15	87	222.8
2	1990	10.35	3.80	184	303.5
2	1991	10.21	3.80	151	206.0
2	1992	13.29	3.80	185	374.5
2	1993	18.58	3.80	262	421.0
2	1994	17.07	3.80	174	595.7
2	1995	16.61	4.25	244	234.8
2	1996	12.28	4.25	159	479.0
2	1997	14.20	3.75	133	170.0