

BASIC IDEAS IN ECONOMETRICS

© Kian-Guan Lim, Singapore Management University, 2001

1. INTRODUCTION

Much of the rigorous statistical work in empirical financial research is based on concepts and methods from the field of econometrics. Some basic ideas of econometrics are presented in this chapter which serves as background preparation for any course in empirical methods of finance at the beginning undergraduate level. Readers with advanced training in statistics or econometrics may skip this chapter.

Financial variables are related one to another, and much of financial research is to discover such relationships so that forecasting and investment decision-making can be done. As a generic example, let two financial variables Y and X be linearly related such that

$$Y = \alpha + \beta X + \varepsilon$$

where α and β represent some specific numerical constants, and ε is a *random variable*.

Definition: A random variable (r.v.) is a numerical quantity at a point in time that can take one of many, possibly infinite number of, values each according to some rule of chance. The set of different possible values that can occur is called the *sample space*. The rule of chance assigns different levels of chances to the values, and is called the *probability density function*. A common example of the latter is the normal probability density function.

The linear relationship of Y and X may be interpreted as follows. At each point in time t , the r.v. ε (pronounced as “ep-si-lon”) by chance assumes a particular value, say -0.03 . At the same time t , X also takes a particular value, say 0.04 . Let α and β be 0.005 and 1.5 respectively. Then the relationship implies that at t , Y must take the value $0.005 + 1.5 \times (0.04) - 0.03$, or 0.035 . You will notice that the randomness in the possible values of ε over time¹ also causes randomness in the values of Y over time. Thus, Y is also a r.v. X may be known at each point in time before Y gets to be revealed or realized; in such a case, X is called a *predetermined variable*. Or, X may itself be a r.v. at time t . In the latter, there will exist a *joint probability density function* between X and ε which will describe the probabilities or chances of values of X and ε jointly or simultaneously occurring. When the nature of X and its

¹ The probability laws of a r.v. over time, or more appropriately a sequence of r.v.’s, are contained in the realm of stochastic processes. We shall see more of this later.

relationship with ε is clearly specified, the linear relationship between Y and X is called a *linear regression*. Y is typically called dependent variable, X the explanatory (or independent) variable, and ε the disturbance, or innovation, or noise.

Many financial models can be represented by such linear regressions. For now, suppose Y is the monthly return rate of stock ABC, and X is the monthly rate of change of the market index. Assume that the joint probability of X and ε is known and that the two r.v.'s are *independent*.

Definition: Two r.v.'s are independent if the probability of any one r.v. taking a particular value is not affected by whatever value the other r.v. takes.

An interesting and relevant question is: given a linear regression model of Y on X, what are the values of alpha α and beta β ? These values are not known in practice, and have to be estimated from available data. In this context, you will note that a high beta implies that the stock ABC's monthly return rate fluctuates with great sensitivity with respect to rate changes in the market index.

Since the relationship between Y and X is a probabilistic one, as opposed to a deterministic one in which knowing X alone would be sufficient to determine Y, it is generally not possible to determine the true values of alpha and beta. However, it is possible to obtain accurate estimates of their true values. The estimates are derived as values to some formulae dependent on realized values of Y and X over time. These formulae are called estimators. These estimators will produce different estimates depending on the different realized values of Y and X over time. Therefore, the estimators themselves are r.v.'s. It will be realized by now that many of the quantities to be estimated are affected by the randomness in the observed values of Y and X.

This chapter commences with a classic two-variable linear regression to motivate an understanding of how financial variables change with reference one to another. It is seen that any further empirical analyses of the regression depend on realized values or outcomes of the r.v.'s Y and X. These outcomes follow certain possibilities described by the probability distributions of Y and X. Therefore, it is crucial to know a bit more about the properties of r.v.'s in general and about the probability distributions describing the randomness. The next three sections deal with these subjects.

2. PROBABILITY DISTRIBUTION

A random variable Y is assumed to have a probability density function (pdf) describing the chances of occurrence of the different possible values of Y . When Y can assume only discrete values, i.e. its sample space is discrete, then the density function is simply called the probability of occurrence of the particular discrete value. For example, the following is a discrete probability distribution.

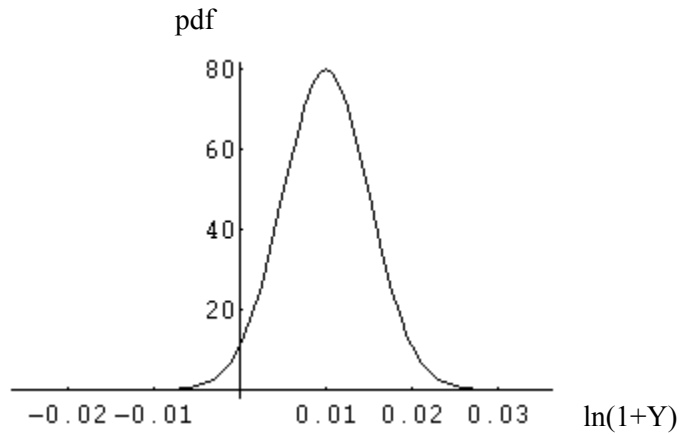
Table 1

Possible Values of Y	Probability
0.035	0.3
0.040	0.5
0.045	0.2

The sample space of Y is $\{0.035, 0.040, 0.045\}$, i.e. comprises three discrete sample values. The value 0.035 occurs with a probability of 0.3, i.e. a chance of 30% or 3 times out of ten, informally speaking. Note that the sum of the probabilities equals one. By convention, the total probability or chance must equal to one or 100%. Often, the r.v. Y can assume any value in a continuous range from say -1 up to infinity ($+\infty$). When the sample space is continuous, Y is said to have a continuous probability distribution. In this case, the pdf of Y is described by a continuous curve, and the sample space of Y can be written as $[-1, +\infty)$. If we take the natural logarithm of $(1+Y)$, then the latter r.v. $\ln(1+Y)$ will have the continuous sample space $(-\infty, +\infty)$.

If Y is the return rate, then the r.v. $\ln(1+Y)$ is the continuously compounded (c.c.) return rate. Assume the c.c. return rate has a normal pdf as follows.

Figure 1



The normal pdf curve has a distinct bell-shape that is symmetrical and drops smoothly on both sides to $-\infty$ and $+\infty$ respectively. It is clear why the original Y cannot have a normal distribution since a normal distribution has supports at $(-\infty, +\infty)$, i.e. the range of the continuous sample space. Therefore natural logarithms are sometimes used in the transform of the original return rates. In practice, the magnitude of the difference between the monthly return rate and the c.c. return rate is small. For economy of notation, we shall continue to employ Y and X except that they represent c.c. return rates instead.

In Figure 1, the continuous pdf curve is used to compute probabilities. For example, the probability of Y taking values in the interval (a,b) is the relative area under the pdf curve within the interval (a,b) . The actual pdf values are by themselves meaningless. Likewise, it is meaningless to measure the probability of Y taking a point value e.g. 0.034. Indeed the latter has a probability measure of zero. In practice the area under a continuous pdf curve is found by integration. The probability of continuous r.v. Y is conveniently expressed in terms of the *cumulative probability function* (cpf). This is also called the probability distribution function. Tables of the cpf for most common probability distributions, including that of the normal distribution, are readily available in most statistics textbooks.

Definition: The cpf of r.v. Y is a mapping which assigns a probability measure in $[0,1]$ to a value y that denotes the event $(-\infty, y)$. The probability of Y taking values in (a,b) is equal to $\text{cpf}(b) - \text{cpf}(a)$.

The probabilistic behavior of r.v. Y is completely characterized by the cpf. In particular, there are several measures of location and variability in the outcomes of Y that are important. These are the *mean*, *standard deviation*, *skewness*, and *kurtosis*. The last two measures are usually less important than the initial two.

Definitions: The mean of a r.v. is the weighted average of its outcomes, using pdf as the weight. The standard deviation of a r.v. is the square root of the weighted average of the square of deviation of outcomes from the mean, using pdf as weight. The square of standard deviation is called variance.

Some examples will make the above definitions clear. Consider the discrete probability distribution given in Table 1 earlier. Since pdf is equivalent to probability itself in discrete distribution, the weights to the outcomes of 0.035, 0.040 and 0.045 are 0.3, 0.5 and 0.2 respectively. Then the mean of Y is

$$0.3(0.035)+0.5(0.040)+0.2(0.045) = 0.0395.$$

The standard deviation of Y is the square root of

$$0.3(0.035-0.0395)^2 + 0.5(0.040-0.0395)^2 + 0.2(0.045-0.0395)^2,$$

or 0.0035. The variance of Y is the square of 0.0035, or 1.225×10^{-5} .

For the continuous probability distribution in Figure 1, the mean is 0.01 and the standard deviation is 0.005. In general, the mean of Y is found by integrating over the product of Y and its pdf. The variance of Y is found by integrating over the product of square of deviation of Y from its mean, and its pdf.

The mean is a measure of central tendency of the r.v., i.e. the value that is the center of gravity of the distribution. The standard deviation and variance are measures of the degree of spread or variability in the r.v. The larger these numbers mean that it is more difficult to predict exactly what the outcome will be. It is important to note that the units of measure of mean and standard deviation are the same as that of the underlying r.v. If the r.v. is a quantity in US\$, then the mean and standard deviation are also measures in US\$. The variance does not have the same unit of measure. Therefore, it is more intuitive to compare variability using standard deviation instead of variance.

Mean and variance are also called the first central moment and the second central moment of the r.v. respectively. Skewness and kurtosis are the third and the fourth central moments of the r.v. The

skewness of a r.v. is the cube root of the weighted average of the cube of deviation of outcomes from the mean, using pdf as weight. The kurtosis of a r.v. is the fourth root of the weighted average of the fourth power of deviation of outcomes from the mean, using pdf as weight. If a pdf curve is symmetrical, then its skewness is zero. If the curve has a longer right tail, the distribution is said to be right-skewed. On the other hand, a longer left tail implies a left-skewed pdf. Higher kurtosis implies a bulky middle part in the shape of the pdf curve, and vice-versa. The skewness of a normal r.v. is zero. The kurtosis of a normal r.v. is three times the square of its variance. These two facts are often used to verify if a r.v. is indeed normal.

3. SAMPLING DISTRIBUTION

In empirical research, it is usually not possible to know the true underlying probability distribution of the financial r.v. in question. Thus it is not possible to compute the mean, standard deviation, skewness, and other parameters of the true distribution. In statistics literature, these parameters are called population parameters. However, estimates of these population parameters can be obtained using the data that are the sample outcomes of the r.v.

If we suppose that the monthly (c.c.) return rate of stock XYZ is a normal r.v. Y with unknown mean μ and unknown standard deviation σ , then reasonable estimates of μ and σ are

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^T y_t$$

and

$$\hat{\sigma} = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (y_t - \hat{\mu})^2}$$

where the collection of values $\{y_1, y_2, y_3, \dots, y_{T-1}, y_T\}$ are a sample of T number of consecutive past monthly return rates that were realized. The subscripts attached to the realized values y denote the association with a particular month. Given these realized values of Y , the estimates $\hat{\mu}$ and $\hat{\sigma}$ are fixed numbers. They may be close to the true parametric values of μ and σ , but will not be exactly equal to them. If we replace realized value y_t by the r.v. Y_t in the above two formulae, then $\hat{\mu}$ and $\hat{\sigma}$ become estimators, not estimates. Estimators are r.v.'s themselves, and thus possess their own probability distributions.

The probability distributions of estimators are called sampling distributions. It can be verified that the means of the sampling distributions of $\hat{\mu}$ and of $\hat{\sigma}$ are μ and σ . This is a desirable property required of estimators so that any realization of estimate will tend to be close to the true parameters. Estimators with this property are called unbiased estimators.

When the sample size T is large, e.g. 100, the standard deviation estimator may be replaced by

$$\hat{\sigma} = \sqrt{\frac{1}{T} \sum_{t=1}^T (y_t - \hat{\mu})^2} .$$

Together with the mean estimator, this latter estimator has the property that its sampling distribution has the least variance. In other words, any realization of estimates will probably be close to the true parameters as chances of large deviations are slim.

Two very interesting statistical findings about sampling distributions were discovered many years ago. The first is called the *Law of Large Numbers*. It is akin to the principle of averaging used in insurance. The second more significant one is called the *Central Limit Theorem*. This is probably the most significant result in the realm of statistical sciences.

Definition: The Law of Large Numbers (LLN) predicts that as the sample size T increases toward infinity, the sample estimates will converge toward the true parameters.

Definition: The Central Limit Theorem (CLT) states that for a large sample size, the estimator $\hat{\mu}$ can be approximated by the normal r.v. The approximation becomes better as the sample size N increases.

For the CLT, the result is ingenious and surprising because the original r.v. Y, from which the sample is drawn, can have any reasonable distribution. Suppose the r.v. Y has mean μ and standard deviation σ . Then the sample mean $\hat{\mu}$ approaches the normal r.v. with mean μ and standard deviation σ / \sqrt{N} . This distributional result is employed extensively for statistical testing.

Although the above two results are obtainable under fairly general conditions, some qualifications are appropriate for mention. As the reader will have noticed, the LLN and the CLT both entail a sequence of r.v.'s. These r.v.'s take realized values which we shall simply call a time series. In statistical terms, the sequence of r.v.'s is called a *stochastic process*. Sample points are to time series as r.v. is to stochastic process.

Definition: A stochastic process is a collection of r.v.'s Y_t such that each Y_t , for every t , has a probability distribution. Furthermore, each Y_t is related to other r.v.'s in a joint multivariate probability distribution.

The important point to note about a stochastic process for our purpose here is that at each time t , there is uncertainty about what is going to happen as embodied in the r.v. Y_t at that time. The joint multivariate probability distribution gives rise to certain features of the process. Generally, the joint probability tells us the chances of what values Y_t will take given that we know what values Y_{t-1} , Y_{t-2} , Y_{t-3} , have taken in the past.

For LLN and CLT to obtain, it is generally required that the stochastic process of Y_t be stationary and ergodic. We can take stationarity to mean that the r.v. Y_t preserves the same distribution over time, i.e. pdf of Y_t is the same as pdf of Y_s for any s different from t . Ergodicity is a technical condition which spells convergence in time series. We shall just mention that a special case of stationarity and ergodicity is the case of identical and independently distributed (i.i.d.) r.v.'s.

4. LINEAR ALGEBRA

Most of the regression techniques that will be covered in this book are best explained through the use of vectors and matrices. The mathematics of these operators is sometimes called linear algebra. It is also possible to use summation operators instead, but the latter are usually tedious and much intuition is also lost in the process. So there is definite gain in learning the new language of linear algebra.

Let \mathbf{x} denote a 3 by 1 vector $\begin{pmatrix} -2 \\ 3 \\ 1 \end{pmatrix}$. Its transpose \mathbf{x}' obtained by putting the n^{th} column as the n^{th} row.

Thus $\mathbf{x}' = (-2 \ 3 \ 1)$. Note that the gap between the numbers facilitates identifying each number within the braces with its position in the vector. Two vectors can be multiplied together to obtain an inner product:

$$\begin{aligned} (-2 \ 3 \ 1) \bullet \begin{pmatrix} 4 \\ -5 \\ 3 \end{pmatrix} &= (-2)(4) + (3)(-5) + (1)(3) \\ &= -20 \end{aligned}$$

which is a scalar here. The inner product operation \bullet is performed by taking the product of the i^{th} element of each vector, and then summing the products. For convenience of exposition, we shall restrict the product to one derived by employing a left row vector and a right column vector.

A matrix is obtained when vectors are joined together, e.g.

$$A = \begin{pmatrix} -2 & 4 \\ 3 & -5 \\ 1 & 3 \end{pmatrix}.$$

In the above, A is a 3 (rows) by 2 (columns) matrix. Just as scalars can be added, subtracted, or multiplied with each other, there are similar matrix operations. For a matrix operation to work between two matrices, they must be conformable in dimensions, as will soon become clear.

Matrix Addition

$$\begin{pmatrix} -2 & 4 \\ 3 & -5 \\ 1 & 3 \end{pmatrix} + \begin{pmatrix} 6 & -3 \\ -5 & 7 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ -2 & 2 \\ 1 & 5 \end{pmatrix}$$

Note that the resulting matrix has the same dimensions as the matrices under addition; each of the ij (i^{th} row, j^{th} column) element of the two matrices is added together to derive the ij element of the resulting matrix. For subtraction, the ij element of the resulting matrix is obtained by the usual scalar subtraction performed on the ij elements of the two matrices on the left-hand side. Matrices are conformable for addition or subtraction provided they have the same dimensions.

Matrix Multiplication

$$\begin{pmatrix} 8 & 2 & -4 \\ 1 & -3 & -5 \end{pmatrix} \cdot \begin{pmatrix} -2 & 4 \\ 3 & -5 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} -14 & 10 \\ -16 & 4 \end{pmatrix}$$

Note that the ij element of the resulting product matrix on the right-hand side is the inner product of the i^{th} row vector of the left matrix and the j^{th} column vector of the right matrix on the left-hand side. Matrices are conformable for multiplication provided the number of columns of the left matrix equals the number of rows of the right matrix. The dimensions of the resulting product matrix are the number of rows of the left matrix and the number of columns of the right matrix.

Special Matrices

The matrix $\mathbf{I}_{n \times n}$ with n rows and n columns and ones along the diagonal but zeros elsewhere is called the identity matrix. For any square matrix (same number of rows and columns) \mathbf{X} , if another matrix can be found such that their product is an identity matrix, then this other matrix is called the *inverse* of \mathbf{X} .

Definition: The inverse of a matrix \mathbf{X} is denoted by \mathbf{X}^{-1} . Furthermore, $\mathbf{X} \bullet \mathbf{X}^{-1} = \mathbf{I}$.

A symmetric matrix is a square matrix such that elements on the upper half of the diagonal are reflections of the elements on the lower half. Formally, \mathbf{A} is symmetric if $\mathbf{A} = \mathbf{A}'$. An *idempotent* matrix is one that when multiplied by itself remains unchanged. This property is useful in constructing test statistics.

Definition: A matrix \mathbf{M} is idempotent if and only if $\mathbf{M} \bullet \mathbf{M} = \mathbf{M}$.

Special Operations

Sometimes special operations are performed on matrices. If a matrix $\mathbf{A}_{n \times m}$ ($n > m$, i.e. more rows than columns) is such that a particular column is equal to a linear combination of the other columns, then \mathbf{A} is said to be *linearly dependent*. An example is the following where the second column is 0.5 times the first column added to 2 times the third column:

$$\begin{pmatrix} 4 & 0 & -1 \\ 2 & -5 & -3 \\ -6 & 7 & 5 \\ 8 & 8 & 2 \\ 2 & 7 & 3 \end{pmatrix}.$$

Definition: A matrix $\mathbf{A}_{n \times m}$ ($n > m$) is linearly independent if no column can be reproduced by any linear combination of the remaining columns. If $n < m$, then it is linearly independent if no row can be reproduced by any linear combination of the remaining rows.

Definition: The rank of a linearly independent matrix $\mathbf{A}_{n \times m}$ ($n > m$) is m .

Definition: The trace (tr) of a square matrix \mathbf{A} is the sum of its diagonal elements. Furthermore, if \mathbf{A} , \mathbf{B} , and \mathbf{C} are matrices such that $\mathbf{A} \bullet \mathbf{B} \bullet \mathbf{C}$ and $\mathbf{C} \bullet \mathbf{A} \bullet \mathbf{B}$ are square matrices, then $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{CAB})$. Note that the matrix product symbol \bullet shall henceforth be skipped for convenience when it is clear from the context that multiplication is involved.

Earlier we see how to compute the mean of a r.v. Y . Whether Y is a discrete or a continuous r.v., we can denote its mean by $E(Y)$, or the expectation of Y . A matrix \mathbf{X} can consist of r.v.'s as its elements.

Likewise we can find the mean of a matrix \mathbf{X} by $E(\mathbf{X})$, which is a matrix such that each element is the mean or expectation of the corresponding element of \mathbf{X} .

5. ORDINARY LEAST SQUARES

Let us get back to where we started, motivating the linear relationship between two r.v.'s Y and X in

$$Y = \alpha + \beta X + \varepsilon$$

where α and β represent some specific numerical constants, and ε is a r.v. We shall also specify that the mean of r.v. ε is zero. This latter assumption is quite innocuous. If the mean of ε were not zero, we can always change its distribution by shifting its mean to zero, then adjusting α to reflect this. The new regression will then explain Y and X exactly as before.

To estimate the parameters α and β , the realized sample data $\{(y_1, x_1), (y_2, x_2), \dots, (y_N, x_N)\}$ are required. Note that the observations y_t and x_t are collected together in pairs since they are observed at the same point in time. Estimation is like trying to locate the being by watching its shadows.

Suppose $\hat{\alpha}$ and $\hat{\beta}$ are the two estimates. If they are good estimates, we shall expect the residuals

$$e_t = y_t - \hat{\alpha} - \hat{\beta}x_t, \quad \text{for every } t,$$

to be close to zero on average. One criterion for this closeness is that the sum of squares of residuals (SSR) \hat{e} is minimized. This is called the Least Squares criterion, and the resulting parameter estimates are called Least Squares Estimates.

Let \mathbf{e} denote vector $(e_1, e_2, \dots, e_N)'$. Let \mathbf{y} denote vector $(y_1, y_2, \dots, y_N)'$. Let \mathbf{b} denote vector $(\hat{\alpha} \quad \hat{\beta})'$. Let \mathbf{x} denote matrix

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}.$$

In matrix notation, therefore

$$\mathbf{e} = \mathbf{y} - \mathbf{x} \mathbf{b}.$$

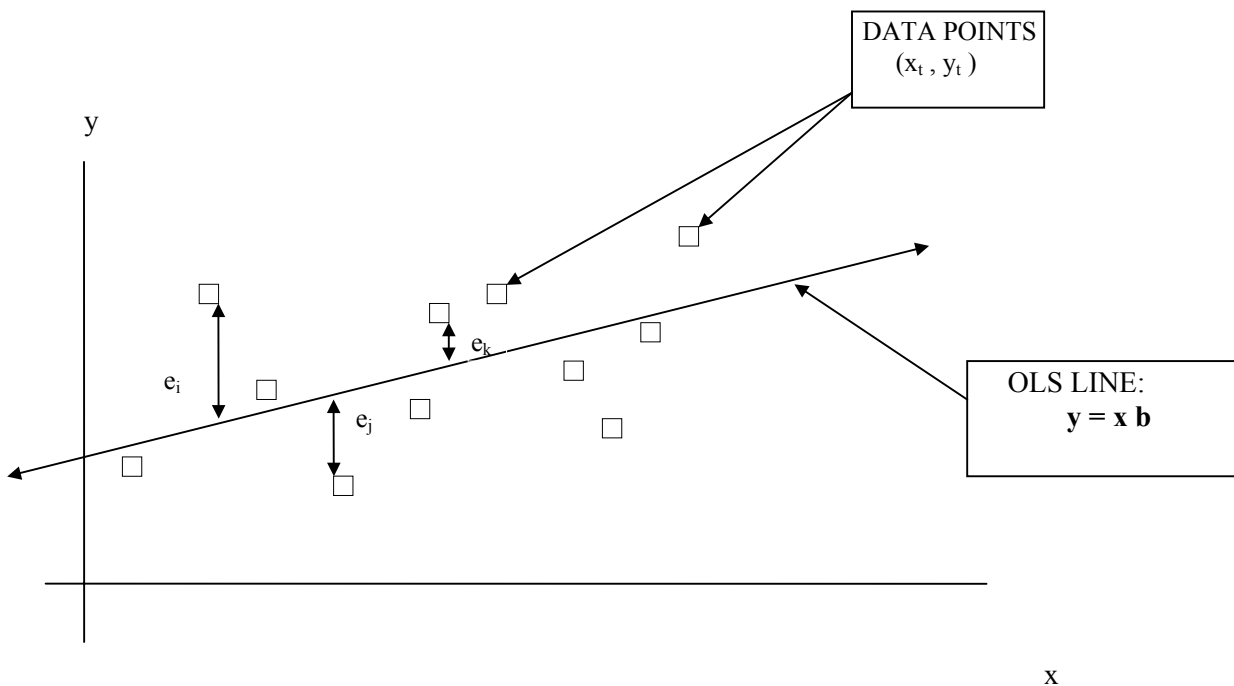
Under Ordinary Least Squares (OLS) estimation, we choose appropriate values of $\hat{\alpha}$ and $\hat{\beta}$ to minimize

$$SSR = \sum_{t=1}^N e_t^2, \text{ or } \mathbf{e}'\mathbf{e}.$$

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{x}\mathbf{b})'(\mathbf{y} - \mathbf{x}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{x}'\mathbf{y} - \mathbf{y}'\mathbf{x}\mathbf{b} + \mathbf{b}'\mathbf{x}'\mathbf{x}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{x}'\mathbf{y} + \mathbf{b}'\mathbf{x}'\mathbf{x}\mathbf{b} \end{aligned}$$

The idea of OLS can be illustrated by the following Figure 2.

FIGURE 2



To minimize $\mathbf{e}'\mathbf{e}$, we partially differentiate the function of \mathbf{b} with respect to \mathbf{b} , i.e. the elements of \mathbf{b} .

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}} = -2\mathbf{x}'\mathbf{y} + 2\mathbf{x}'\mathbf{x}\mathbf{b}$$

For minimum, equate this derivative to zero for the necessary condition; the sufficient condition involving the second order condition is taken care of because the quadratic equation will ensure the minimum:

$$\mathbf{x}'\mathbf{x} \mathbf{b} = \mathbf{x}'\mathbf{y} \quad \Leftrightarrow \quad \mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{y} .$$

The right arrow \Rightarrow means “implies” while the left arrow \Leftarrow means “is implied by”. Thus we find the OLS estimates in \mathbf{b} . If we express \mathbf{b} in summation algebra, it is the following:

$$\hat{\beta} = \frac{\sum_{t=1}^N (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^N (x_t - \bar{x})^2} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} ,$$

where \bar{x} and \bar{y} are the sample means of x and y . Note that $\hat{\beta}$ can also be expressed as

$$\sum_{t=1}^N w_t y_t \quad \text{where} \quad w_t = \frac{(x_t - \bar{x})}{\sum_{t=1}^N (x_t - \bar{x})x_t} .$$

Clearly, the weights w_t do not depend on y_t ; $\hat{\beta}$, thus also $\hat{\alpha}$, are linear estimates, i.e. formed by using linear combinations of sample $\{y_t\}$.

Suppose we ask the question about the variability of the estimator which gives rise to estimate \mathbf{b} . It may be relevant to ask this way: given that we observe the data \mathbf{x} , how variable is the estimator \mathbf{B} ? Notice that \mathbf{B} is also dependent on r.v. \mathbf{Y} ; therefore, it is a r.v. More precisely, \mathbf{B} is a *conditional* r.v.

Definition: A r.v. Y is said to be conditional on observation x if the pdf of Y is obtained by replacing terms in the pdf involving X with x . The resulting pdf is called the conditional pdf.

Another way to motivate conditional r.v. is to think of \mathbf{x} above as predetermined variables, which are to be treated as if they are constants. In such case, the *covariance* between x_t and ε_t is zero. Zero covariance is implied by independence between x_t and ε_t , though it is weaker than independence, i.e. it does not turn implies independence. Covariance expresses a statistical relationship between two r.v.'s.

Definition: The covariance of two r.v.'s Y and X is the mean of the product of the deviation of each r.v. from its own mean. The covariance can be expressed as

$$\text{Cov}(Y, X) = E((Y - \mu_Y)(X - \mu_X))$$

where $\mu_Y = E(Y)$ and $\mu_X = E(X)$. The covariance of a r.v. with itself is its variance.

It is seen that if Y tends to be high when X is high, and vice-versa, then $\text{Cov}(Y, X)$ is positive. If Y tends to be low when X is high, and vice-versa, then $\text{Cov}(Y, X)$ is negative. Covariance measures the tendency for the two r.v.'s to either move together in the same direction, or in opposite direction. If the covariance is zero, then it implies that the two r.v.'s tend to move each its own way without being “affected” by the other. The existence of covariance measures necessitates the presence of a joint probability distribution of the r.v.'s Y and X. With two r.v.'s, this joint distribution is called a bivariate distribution. We shall return to this topic later in a later chapter.

Consider the estimator $\mathbf{B} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{Y}$ where we replace the sample values \mathbf{y} by their corresponding r.v.'s \mathbf{Y} . Then \mathbf{B} itself is a random vector, i.e. a vector of r.v.'s. The mean and variance of the sampling distribution of \mathbf{B} is found as follows.

$$\begin{aligned} \mathbf{B} &= (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{Y} \\ &= (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'(\mathbf{x} \mathbf{b}_0 + \mathbf{E}) \\ &= (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{x} \mathbf{b}_0 + (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{E} \\ &= \mathbf{b}_0 + (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{E} \end{aligned}$$

where \mathbf{E} denotes the N by 1 vector of r.v.'s ε_i , and \mathbf{b}_0 denotes the vector of true parameters $(\alpha \ \beta)'$.

$$\begin{aligned} E(\mathbf{B}) &= \mathbf{b}_0 + (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' E(\mathbf{E}) \\ &= \mathbf{b}_0 \end{aligned}$$

since $E(\mathbf{E}) = 0$. Thus we see that OLS estimator \mathbf{B} is unbiased, i.e. the sampling distribution of \mathbf{B} is centered with mean at the true parametric vector \mathbf{b}_0 .

In considering the variance of \mathbf{B} , we first have to define what a covariance matrix is. Let the covariance matrix of the r.v.'s $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{N-1}, \varepsilon_N$ be Ω , a N by N matrix.

$$\mathbf{\Omega}_{N \times N} = \begin{pmatrix} \text{Cov}(\varepsilon_1, \varepsilon_1) & \text{Cov}(\varepsilon_1, \varepsilon_2) & \cdots & \cdots & \text{Cov}(\varepsilon_1, \varepsilon_N) \\ \text{Cov}(\varepsilon_2, \varepsilon_1) & \text{Cov}(\varepsilon_2, \varepsilon_2) & \cdots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\varepsilon_N, \varepsilon_1) & \cdots & \cdots & \cdots & \text{Cov}(\varepsilon_N, \varepsilon_N) \end{pmatrix}$$

The above is also called the covariance matrix of vector \mathbf{E} . Notice that the Covariance matrix is a symmetrical matrix since the ij element is equal to the ji element. We shall assume that the stochastic process \mathbf{E} is i.i.d. Therefore the diagonal terms in $\mathbf{\Omega}$ are all equal to $\text{Var}(\varepsilon)$, a constant, say σ_E^2 . The off-diagonal terms are all equal to zero. Now,

$$\begin{aligned} \text{Var}(\mathbf{B}) &= E[(\mathbf{B} - \mathbf{b}_0)(\mathbf{B} - \mathbf{b}_0)'] \\ &= E[(\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{E} \mathbf{E}' \mathbf{x} (\mathbf{x}'\mathbf{x})^{-1}] \\ &= (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' E(\mathbf{E} \mathbf{E}') \mathbf{x} (\mathbf{x}'\mathbf{x})^{-1} \\ &= (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' \sigma_E^2 \mathbf{I}_N \mathbf{x} (\mathbf{x}'\mathbf{x})^{-1} \\ &= \sigma_E^2 (\mathbf{x}'\mathbf{x})^{-1} . \end{aligned}$$

The variance of two by one vector \mathbf{B} , $\text{Var}(\mathbf{B})$, is a two by two covariance matrix.

Theorem: The Gauss-Markov theorem states that the OLS estimator \mathbf{B} has the smallest variance among all linear unbiased estimators of \mathbf{b}_0 . Thus we say that \mathbf{B} is BLUE, i.e. best linear unbiased estimator.

The OLS technique and results can be easily extended to multivariate model, i.e. more than two variables. For example, there could be four explanatory variables for Y , in the form of r.v.'s U, V, W , and X , and including also the constant. Then the sample of \mathbf{x} would be

$$x = \begin{pmatrix} 1 & u_1 & v_1 & w_1 & x_1 \\ 1 & u_2 & v_2 & w_2 & x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & u_N & v_N & w_N & x_N \end{pmatrix} .$$

The parameters would be

$$b = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} .$$

The OLS regression equation would still be written as $\mathbf{y} = \mathbf{x} \mathbf{b} + \mathbf{e}$. Estimates of $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ would similarly be estimated by $\mathbf{b} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{y}$.

6. GENERALIZED LEAST SQUARES

In the last section, it is assumed that the disturbance ε has covariance matrix $\sigma_E^2 \mathbf{I}_N$. We call the disturbance homoskedastic, i.e. “homogeneous”. Sometimes it is also referred to as spherical disturbance. However, it is possible that the N by N covariance matrix of the disturbance ε is not homoskedastic, in which case it is called heteroskedastic. Since covariance matrix by nature must be *positive definite*, let the covariance matrix of ε be represented by symmetrical matrix $\sigma_E^2 \mathbf{\Omega}$ where in general the off-diagonal elements are non-zero.

Definition: A N by N square matrix $\mathbf{\Omega}$ is called positive definite if and only if (\Leftrightarrow) any conformable real N by 1 vector \mathbf{w} yields $\mathbf{w}' \mathbf{\Omega} \mathbf{w}$ as a positive real number, i.e. ≥ 0 .

Therefore, $E(\mathbf{E}\mathbf{E}') = \sigma_E^2 \mathbf{\Omega}$. If $\mathbf{\Omega}$ is known, its positive definiteness implies that there exists a *non-singular* N by N matrix \mathbf{p} such that $\mathbf{\Omega} = \mathbf{p} \mathbf{p}'$.

Definition: A non-singular matrix is one in which its inverse exists. A singular matrix is one without an inverse, i.e. has zero determinant.

Therefore we can take the inverse of \mathbf{p} on the left side, and the inverse of \mathbf{p}' on the right side, and obtain $\mathbf{p}^{-1} \mathbf{\Omega} \mathbf{p}^{-1} = \mathbf{I}$. Also, $\mathbf{p}^{-1} \mathbf{p}' = \mathbf{\Omega}^{-1}$.

The question at hand is how the heteroskedasticity in the disturbance covariance matrix will affect the OLS estimator. The OLS estimator in this case is still unbiased. However, it loses the property of minimum variance. It is also incorrect now to express the covariance matrix of the OLS estimator in the way done earlier. With heteroskedasticity, a BLUE can still be obtained in the following way.

Transform the r.v. \mathbf{Y} and \mathbf{X} as follows. $\mathbf{Z} = \mathbf{p}^{-1} \mathbf{Y}$; $\mathbf{W} = \mathbf{p}^{-1} \mathbf{X}$ and $\mathbf{U} = \mathbf{p}^{-1} \mathbf{E}$. Pre-multiply the original model $\mathbf{Y} = \mathbf{X} \mathbf{b}_0 + \mathbf{E}$ (where it should be remembered that \mathbf{E} is now heteroskedastic) by \mathbf{p}^{-1} . So we obtain

$$\mathbf{Z} = \mathbf{W} \mathbf{b}_0 + \mathbf{U}.$$

The covariance matrix of \mathbf{U} is

$$\begin{aligned} E(\mathbf{U}\mathbf{U}') &= E(\mathbf{p}^{-1} \mathbf{E} \mathbf{E}' \mathbf{p}^{-1}) \\ &= \sigma_E^2 \mathbf{p}^{-1} \mathbf{\Omega} \mathbf{p}^{-1}, \\ &= \sigma_E^2 \mathbf{I}_N. \end{aligned}$$

Thus the transformed regression now has homoskedastic disturbance u_t . To obtain BLUE, apply the same OLS to the transformed regression:

$$\begin{aligned} \mathbf{b} &= (\mathbf{w}'\mathbf{w})^{-1} \mathbf{w}' \mathbf{z} \\ &= (\mathbf{x}' \mathbf{p}^{-1} \mathbf{p}^{-1} \mathbf{x})^{-1} \mathbf{x}' \mathbf{p}^{-1} \mathbf{p}^{-1} \mathbf{y} \\ &= (\mathbf{x}' \mathbf{\Omega}^{-1} \mathbf{x})^{-1} \mathbf{x}' \mathbf{\Omega}^{-1} \mathbf{y}. \end{aligned}$$

This is called the Generalized Least Squares (GLS) estimate of the original regression involving \mathbf{y} and \mathbf{x} .

$$\begin{aligned}\text{Var}(\mathbf{b}) &= \sigma_E^2 (\mathbf{w}'\mathbf{w})^{-1} \\ &= \sigma_E^2 (\mathbf{x}' \boldsymbol{\Omega}^{-1} \mathbf{x})^{-1} .\end{aligned}$$

Further References

There are some excellent books written in the subject area of econometrics which are worth buying.

Jack Johnston John Dinardo: [Econometric Methods](#), McGraw Hill 1996.

Fumio Hayashi, *Econometrics*, Princeton University Press, 2000.

Greene W., *Econometric Analysis*, 2nd Edition, Prentice Hall, NJ, 1993.

Damodar N. Gujarati, *Basic Econometrics*, McGraw-Hill, New York, 1995.

Jan Kmenta, *Elements of Econometrics*, University of Michigan Press, 1997.