LECTURE 1

Conditional Expectations and Regression Analysis

In this chapter, we shall study three methods that are capable of generating estimates of statistical parameters in a wide variety of contexts. These are the method of moments, the method of least squares and the principle of maximum likelihood.

The methods will be studied only in relation to the simple linear regression model; and it will be seen that each entails assumptions that may be more or less appropriate to the context in which the model is to be applied.

In the case of the regression model, the three methods generate estimating equations that are formally identical; but this does not justify us in taking a casual approach to the statistical assumptions that sustain the model. To be casual in making our assumptions is to invite the danger of misinterpretation when the results of the estimation are in hand.

We begin with the method of moments, we shall proceed to the method of least squares, and we shall conclude with a brief treatment of the method of maximum likelihood.

Conditional Expectations

Let y be a continuously distributed random variable whose probability density function is $f(y)$. If we wish to predict the value of y without the help of any other information, then we might take its expected value, which is defined by

$$E(y) = \int y f(y) dy.$$

The expected value is a so-called minimum-mean-square-error (m.m.s.e.) predictor. If π is the value of a prediction, then the mean-square error is given by

$$\begin{aligned} M &= \int (y-\pi)^2 f(y) dy \\ &= E\big\{(y-\pi)^2\big\} \\ &= E(y^2) - 2\pi E(y) + \pi^2; \end{aligned} \tag{1}$$

and, using the methods of calculus, it is easy to show that this quantity is minimised by taking $\pi = E(y)$.

Now let us imagine that y is statistically related to another random variable x, whose values have already been observed. For the sake of argument, it may be assumed that the form of the joint distribution of x and y, which is $f(x,y)$, is known. Then, the minimum-mean-square-error prediction of y is given by the conditional expectation

$$E(y|x) = \int y \frac{f(x,y)}{f(x)} dy \tag{2}$$

wherein

$$f(x) = \int f(x,y) dy \tag{3}$$

is the so-called marginal distribution of x. This proposition may be stated formally in a way that will assist us in proving it:

(4) Let $\hat{y} = \hat{y}(x)$ be the conditional expectation of y given x, which is also expressed as $\hat{y} = E(y|x)$. Then $E\{(y-\hat{y})^2\} \leq E\{(y-\pi)^2\}$, where $\pi = \pi(x)$ is any other function of x.

Proof. Consider

$$\begin{aligned} E\big\{(y-\pi)^2\big\} &= E\Big[\big\{(y-\hat{y}) + (\hat{y}-\pi)\big\}^2\Big] \\ &= E\big\{(y-\hat{y})^2\big\} + 2E\big\{(y-\hat{y})(\hat{y}-\pi)\big\} + E\big\{(\hat{y}-\pi)^2\big\}. \end{aligned} \tag{5}$$

In the second term, there is

$$\begin{aligned} E\big\{(y-\hat{y})(\hat{y}-\pi)\big\} &= \int_x \int_y (y-\hat{y})(\hat{y}-\pi) f(x,y) \partial y \partial x \\ &= \int_x \left\{ \int_y (y-\hat{y}) f(y|x) \partial y \right\} (\hat{y}-\pi) f(x) \partial x \\ &= 0. \end{aligned} \tag{6}$$

Here, the second equality depends upon the factorisation $f(x,y) = f(y|x)f(x)$, which expresses the joint probability density function of x and y as the product of the conditional density function of y given x and the marginal density function of x. The final equality depends upon the fact that $\int (y-\hat{y}) f(y|x) \partial y = E(y|x) - E(y|x) = 0$. Therefore, $E\{(y-\pi)^2\} = E\{(y-\hat{y})^2\} + E\{(\hat{y}-\pi)^2\} \geq E\{(y-\hat{y})^2\}$, and the assertion is proved.

The definition of the conditional expectation implies that

$$\begin{aligned} E(xy) &= \int_x \int_y xyf(x,y)\partial y\partial x \\ &= \int_x x\left\{ \int_y yf(y|x)\partial y\right\} f(x)\partial x \\ &= E(x\hat{y}). \end{aligned} \tag{7}$$

When the equation $E(xy) = E(x\hat{y})$ is rewritten as

$$E\big\{x(y-\hat{y})\big\} = 0, \tag{8}$$

it may be described as an orthogonality condition. This condition indicates that the prediction error $y - \hat{y}$ is uncorrelated with x. The result is intuitively appealing; for, if the error were correlated with x, then the information of x could not have been used efficiently in forming $\hat{y}$.

If the joint distribution of x and y is a normal distribution, then it is straightforward to find an expression for the function $E(y|x)$. In the case of a normal distribution, there is

$$E(y|x) = \alpha + \beta x, \tag{9}$$

which is to say that the conditional expectation of y given x is a linear function of x. Equation (9) is described as a linear regression equation; and this terminology will be explained later.

The object is to find expressions for α and β that are in terms of the first-order and second-order moments of the joint distribution. That is to say, we wish to express α and β in terms of the expectations $E(x)$, $E(y)$, the variances $V(x)$, $V(y)$ and the covariance $C(x,y)$.

Admittedly, if we had already pursued the theory of the Normal distribution to the extent of demonstrating that the regression equation is a linear equation, then we should have already discovered these expressions for α and β. However, present purposes are best served by taking equation (9) as the starting point; and the linearity of the regression equation may be regarded as an assumption in its own right rather than as a deduction from the assumption of a normal distribution.

To begin, equation (9) may be multiplying throughout by $f(x)$, and integrates with respect to x. This gives

$$E(y) = \alpha + \beta E(x), \tag{10}$$

whence

$$\alpha = E(y) - \beta E(x). \tag{11}$$

Equation (10) shows that the regression line passes through the point $E(x,y) = \{E(x), E(y)\}$, which is the expected value of the joint distribution.

Putting (11) into (9) gives

$$E(y|x) = E(y) + \beta\big\{x - E(x)\big\}, \tag{12}$$

which shows that the conditional expectation of y differs from the unconditional expectation in proportion to the error of predicting x by taking its expected value.

Next, (9) is multiplied by x and $f(x)$ and then integrated with respect to x to provide

$$E(xy) = \alpha E(x) + \beta E(x^2). \tag{13}$$

Multiplying (10) by $E(x)$ gives

$$E(x)E(y) = \alpha E(x) + \beta\big\{E(x)\big\}^2, \tag{14}$$

whence, on taking (14) from (13), we get

$$E(xy) - E(x)E(y) = \beta\Big[E(x^2) - \big\{E(x)\big\}^2\Big], \tag{15}$$

which implies that

$$\begin{aligned}\beta &= \frac{E(xy) - E(x)E(y)}{E(x^2) - \big\{E(x)\big\}^2}\\ &= \frac{E\Big[\big\{x - E(x)\big\}\big\{y - E(y)\big\}\Big]}{E\Big[\big\{x - E(x)\big\}^2\Big]}\\ &= \frac{C(x,y)}{V(x)}.\end{aligned} \tag{16}$$

Thus, α and β have been expressed in terms of the moments $E(x)$, $E(y)$, $V(x)$ and $C(x,y)$ of the joint distribution of x and y.

Example. Let $x = \xi + \eta$ be an observed random variable which combines a signal component ξ and a noise component η. Imagine that the two components are uncorrelated with $C(\xi,\eta) = 0$, and let $V(\xi) = \sigma_\xi^2$ and $V(\eta) = \sigma_\eta^2$. The object is to extract the signal from the observation.

According to the formulae of (12) and (16), the expectation of the signal conditional upon the observation is

$$E(\xi|x) = E(\xi) + \frac{C(x,\xi)}{V(x)}\{x - E(x)\}. \tag{17}$$

Given that ξ and η are uncorrelated, it follows that

$$V(x) = V(\xi + \eta) = \sigma_\xi^2 + \sigma_\eta^2 \tag{18}$$

and that

$$C(x,\xi) = V(\xi) + C(\xi,\eta) = \sigma_\xi^2. \tag{19}$$

Therefore

$$E(\xi|x) = E(\xi) + \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\eta^2}\{x - E(x)\}. \tag{20}$$

Estimation by the Method of Moments

The values of the various moments comprised in the formulae for the regression parameters are unlikely to be know in the absense of sample data. However, they are easily estimated from the data. Imagine that a sample of T observations on x and y is available: $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$. Then, the following empirical or sample moments can be calculated:

$$\begin{aligned}
\bar{x} &= \frac{1}{T}\sum_{t=1}^{T} x_t,\\
\bar{y} &= \frac{1}{T}\sum_{t=1}^{T} y_t,\\
s_x^2 &= \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2 = \frac{1}{T}\sum_{t=1}^{T} x_t^2 - \bar{x}^2,\\
s_{xy} &= \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y}) = \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \bar{x}\bar{y}.
\end{aligned} \tag{21}$$

The method of moments suggests that, in order to estimate α and β, the moments should be replaced in the formulae of (11) and (16) by the corresponding sample moments. Thus the estimates of α and β are

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x},\\
\hat{\beta} &= \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}.
\end{aligned} \tag{22}$$

The justification of the method is that, in many of the circumstances under which the data are liable to be generated, the sample moments are expected to converge to the true moments of the bivariate distribution, thereby causing the estimates of the parameters to converge, likewise, to the true values.

In this context, the concept of convergence has a special definiton. According to the concept of convergence which is used in mathematical analysis,

(23) A sequence of numbers $\{a_n\}$ is said to converge to a limit a if, for any arbitrarily small real number ϵ, there exists a corresponding integer N such that $|a_n - a| < \epsilon$ for all $n \geq N$.

This concept is not appropriate to the case of a stochastic sequence, such as a sequence of estimates. For, no matter how many observations N have been incorporated in the estimate a_N, there remains a possibility that, subsequently, an aberrant observation y_n will draw the estimate a_n beyond the bounds of $a \pm \epsilon$. A criterion of convergence must be adopted that allows for this possibility:

(24) A sequence of random variables $\{a_n\}$ is said to converge weakly in probability to a limit a if, for any ϵ, there is $\lim P(|a_n - a| > \epsilon) = 0$ as $n \rightarrow \infty$ or, equivalently, $\lim P(|a_n - a| \leq \epsilon) = 1$.

This means that, by increasing the size of the sample, we can make it virtually certain that a_n will 'fall within an epsilon of a.' It is conventional to describe a as the probability limit of a_n and to write $\text{plim}(a_n) = a$.

The virtue of this definition of convergence is that it does not presuppose that the random variable a_n has a finite variance or even a finite mean. However, if a_n does have finite moments, then a concept of mean-square convergence can be employed.

(25) A sequence of random variables $\{a_n\}$ is said to converge in mean square to a limit a if $\lim(n \rightarrow \infty) E\{(a_n - a)^2\} = 0$.

It should be noted that

$$\begin{aligned} E\Big\{\big(a_n - a\big)^2\Big\} &= E\Big\{\Big(\big[a_n - E(a_n)\big] - \big[a - E(a_n)\big]\Big)^2\Big\} \\ &= V(a_n) + E\Big[\big\{a - E(a_n)\big\}^2\Big]; \end{aligned} \tag{26}$$

which is to say that the mean-square error of a_n is the sum of its variance and the square of its bias. If a_n is to converge in mean square to a, then both of these quantities must vanish.

Convergence in mean square is a stronger condition than convergence in probability in the sense that it implies the latter. Whenever an estimator converges in probability to the value of the parameter which it purports to represent, then it is said to be a consistent estimator.

Regression and the Eugenic Movement

The theory of linear regression has its origins in the late 19th century when it was closely associated with the name of the English eugenicist Francis Galton (1822–1911).

Galton was concerned with the hereditibility of physical and mental characteristics; and he sought ways of improving the genetic quality of the human race. His disciple Karl Pearson, who espoused the same eugenic principles as Galton and who was a leading figure in the early development of statistical theory in Britain, placed Galton's contributions to science on a par with those of Charles Darwin who was Galton's cousin.

Since the 1930's, the science of eugenics has fallen into universal disrepute, and its close historical association with statistics has been largely forgotten. However it should be recalled that one of the premier journals in its field, which now calls itself the *Annals of Human Genetics*, began life as *The Annals of Eugenics.* The thoughts which inspired the Eugenic Movement still arise, albeit that they are expressed, nowadays, in different guises.

One of Galton's studies that is best remembered concerns the relationship between the heights of fathers and the heights of their sons. The data that was gathered was plotted on a graph and it was found to have a distribution that resembles a bivariate normal distribution.

It might be supposed that the best way to predict the height of a son is to take the height of the father. In fact, such a method would lead of a systematic over-estimation of the height of the sons if their fathers were above-average height. In the terminology of Galton, we usually witness a regression of the son's height towards "mediocrity ".

Galton's terminology suggests a somewhat unbalanced point of view. The phenomenon of regression is accompanied by a corresponding phenomenon of progression whereby fathers of less than average height are liable to have sons who are taller than themselves. Also, if the distribution of heights is to remain roughly the same from generation to generation and if it is not to loose its dispersion, then there are bound to be cases which conflict with the expectation of an overall reversion towards the mean.

A little reflection will go a long way toward explaining the phenomenon of reversion; for we need only consider the influence of the mother's height. If we imagine that, in general, men of above-average height show no marked tendency to marry tall women, then we might be prepared to attribute an average height to the mother, regardless of the father's height. If we acknowledge that the two parents are equally influential in determining the physical characteristics of their offspring, then we have a ready explanation of the tendency of heights

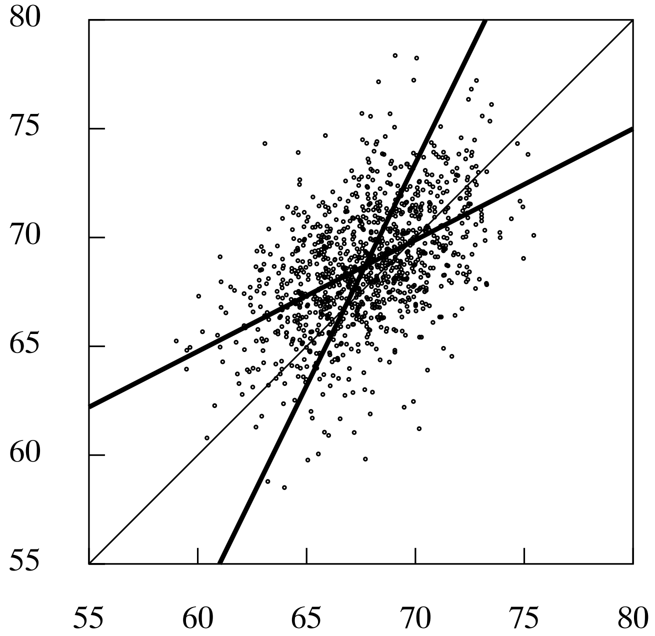

Figure 1. Pearson's data comprising 1078 measurements of on the heights of father (the abscissae) and of their sons (the ordinates), together with the two regression lines. The correlation coefficient is 0.5013.

to revert to the mean. To the extent that tall people choose tall partners, we shall see a retardation of the tendency; and the characteristics of abnormal height will endure through a greater number of generations.

An investigation into the relationship between the heights of parents and the heights of their offspring was published in 1886 by Francis Galton. He collected the data of 928 adult offspring and their parents. He combined the height of the two parents by averaging the father's height and the mother's height scaled by a factor of 1.08, which was obtained by a comparison of the average male height and the average female height. This created a total of 205 midparents. Likewise, all female heights were multiplied by a factor 1.08. Even when the sexes are combined in this manner there, is a clear regression towards the mean.

Galton's analysis was extended by Karl Pearson (1857–1936) in a series of papers. In 1903, Pearson and Lee published an analysis that comprised separate data on fathers and sons and on mothers and daughters. Figure 1 is based on 1078 measurements from Pearson's data of a father's height and his son's height. It appears to indicate that, in the late 19th century, there was a small but significant increase in adult male stature.

The Bivariate Normal Distribution

Most of the results in the theory of regression that have described so far can be obtained by examining the functional form of the bivariate normal distribution. Let x and y be the two variables. Let their means be denoted by

$$E(x) = \mu_x, \qquad E(y) = \mu_y, \tag{27}$$

their variances by

$$V(x) = \sigma_x^2, \qquad V(y) = \sigma_y^2 \tag{28}$$

and their covariance by

$$C(x, y) = \rho\sigma_x\sigma_y. \tag{29}$$

Here

$$\rho = \frac{C(x, y)}{\sqrt{V(x)V(y)}}, \tag{30}$$

which is called the correlation coefficient of x and y, provides a measure of the relatedness of these variables.

The Cauchy–Schwarz inequality indicates that $-1 \leq \rho \leq 1$. If $\rho = 1$, then there is an exact positive linear relationship between the variables whereas, if $\rho = -1$, then there is an exact negative linear relationship. Neither of these extreme cases is admissible in the present context for, as can be seen by examining the following formulae, they lead to the collapse of the bivariate distribution.

The bivariate distribution is specified by

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp Q(x, y), \tag{31}$$

where

$$Q = \frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\} \tag{32}$$

is a quadratic function of x and y.

The function can also be written as

$$Q = \frac{-1}{2(1-\rho^2)}\left\{\left(\frac{y-\mu_y}{\sigma_y} - \rho\frac{x-\mu_x}{\sigma_x}\right)^2 - (1-\rho^2)\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right\}. \tag{33}$$

Thus, there is

$$f(x, y) = f(y|x)f(x), \tag{34}$$

where

$$f(x)=\frac{1}{\sigma_x\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}, \tag{35}$$

and

$$f(y|x)=\frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}}\exp\left\{-\frac{(y-\mu_{y|x})^2}{2\sigma_y^2(1-\rho)^2}\right\}, \tag{36}$$

with

$$\mu_{y|x}=\mu_y+\frac{\rho\sigma_y}{\sigma_x}(x-\mu_x). \tag{37}$$

Equation (37) is the linear regression equation, which specifies the value of $E(y|x)=\mu_{y|x}$ in terms of x; and it is simply the equation (12) in another notation. Equation (36) indicates that the variance of y about its conditional expectation is

$$V(y|x)=\sigma_y^2(1-\rho^2). \tag{38}$$

Since $(1-\rho^2)\leq 1$, it follows that variance of the conditional predictor $E(y|x)$ is less than that of the unconditional predictor $E(y)$ whenever $\rho\neq 0$—which is whenever there is a correlation between x and y. Moreover, as this correlation increases, the variance of the conditional predictor diminishes.

There is, of course, a perfect symmetry between the arguments x and y in the bivariate distribution. Thus, if we choose to factorise the joint probability density function as $f(x,y)=f(x|y)f(y)$, then, to obtain the relevant results, we need only interchange the x's and the y's in the formulae above.

The fact shuould be noted that x and y will be statistically independent random variables that are uncorrelated with each other if and only if their joint distribution can be factorised as the product of their marginal distributions: $f(y,y)=f(y)f(y)$. In the absence of statistical independence, the joint distribution becomes the product of a conditional distribution and a marginal distribution: $f(y,x)=f(y|x)f(x)$. The arguments of these two distributions will retain the properties of statistical independence. That is to say, the random variables $\varepsilon=y-\mu_{y|x}$ and $\nu=x-\mu_x$ are, by construction, statistically independent with $C(\varepsilon,\nu)=0$.

Least-Squares Regression Analysis

Galton's analysis, which described the regression relationship between the heights of fathers and their sons, was an exercise in descriptive statistics based on a given set of data. There can be no presumption that, for a different race of men living in a different environment, the same parameters would be uncovered. It is only as an experiment in thought that we may vary the value of the explanatory variable x and watch the concomitant variation of y. The heights of individual men are not subject to experimental manipulation.

Econometrics, in contrast to descriptive statistics, is often concerned with functional regression relationships, which purport to describe the effects of manipulating the instruments of economic policy such as interest rates and rates of taxation. In such cases, it is no longer appropriate to attribute a statistical distribution to the explanatory variable x, which now assumes the status of a control variable. Therefore, it is necessary to derive the formulae of regression analysis from principles that make no reference to the joint distribution of the variables. The principle of least squares is appropriate to this purpose.

Before admitting this change of emphasis, we should offer some words of caution. For it seems that many of the errors of applied econometrics arise when an analyst imagines that, in fitting a regression equation, he has uncovered a causal connection.

The data that are used in inferring a regression relationship are part of an historical record of the evolution of the economy; and it is never certain that the same statistical relationships would have prevailed in other circumstances. Nor is it clear that they will prevail in the future.

An econometric analysis is often conducted with a view to guiding the actions of a regulatory agent. However, such actions are liable to alter the statistical relationships prevailing amongst economic variables. An assertion that a particular relationship will endure through time and that it will be unaffected by regulatory intercessions ought to be greeted with skepticism. Yet, in such matters, applied econometricians are often eager to suspend their disbelief.

To assist the application of the method of least squares, the regression equation, which has been defined by $E(y|x) = \alpha + \beta x$, can be written, alternatively, as

$$y = \alpha + x\beta + \varepsilon, \tag{39}$$

where $\varepsilon = y - E(y|x)$ is a random variable with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$. This equation may be used to depict a functional relationship between an independent variable x and a dependent variable y. The relationship is affected by a disturbance ε, which is independent of x and which might be taken to represent the effect of a large number of variables of minor importance that are not taken into account explicitly in describing the relationship.

Imagine that there is a sample of observations $(x_1, y_1), \ldots, (x_T, y_T)$ and that, from these data, the parameters α and β are to be estimated. The principle of least squares suggests that this should be done by choosing the values that minimise the quantity

$$\begin{aligned} S &= \sum_{t=1}^{T} \varepsilon_t^2 \\ &= \sum_{t=1}^{T} (y_t - \alpha - x_t\beta)^2. \end{aligned} \tag{40}$$

This is the sum of squares of the vertical distances—measured parallel to the y-axis—of the data points from an interpolated regression line.

Differentiating the function S with respect to α and setting the results to zero for a minimum gives

$$\begin{aligned} -2\sum(y_t - \alpha - \beta x_t) = 0, &\quad \text{or, equivalently,} \\ \bar{y} - \alpha - \beta\bar{x} = 0. & \end{aligned} \tag{41}$$

This generates the following estimating equation for α:

$$\alpha(\beta) = \bar{y} - \beta\bar{x}. \tag{42}$$

Next, differentiating with respect to β and setting the result to zero gives

$$-2\sum x_t(y_t - \alpha - \beta x_t) = 0. \tag{43}$$

On substituting for α from (42) and eliminating the factor -2, this becomes

$$\sum x_t y_t - \sum x_t(\bar{y} - \beta\bar{x}) - \beta\sum x_t^2 = 0, \tag{44}$$

whence we get

$$\begin{aligned} \hat{\beta} &= \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2} \\ &= \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}. \end{aligned} \tag{45}$$

This expression is identical to the one under (22) which has been derived via the method of moments. Putting $\hat{\beta}$ into the estimating equation for α under (42) gives the same estimate $\hat{\alpha}$ for the intercept parameter as the one to be found under (22).

The method of least squares does not automatically provide an estimate of $\sigma^2 = E(\varepsilon_t^2)$. To obtain an estimate, the method of moments is invoked. In view of the fact that the regression residuals $e_t = y_t - \hat{\alpha} - \hat{\beta}x_t$ represent estimates of the corresponding values of ε_t, the method suggests the following estimator:

$$\tilde{\sigma}^2 = \frac{1}{T}\sum e_t^2. \tag{46}$$

In fact, this is a biased estimator with

$$E\left(\frac{\tilde{\sigma}^2}{T}\right) = \left(\frac{T-2}{T}\right)\sigma^2; \tag{47}$$

so it is common to adopt the unbiased estimator

$$\hat{\sigma}^2 = \frac{\sum e_t^2}{T-2}. \tag{48}$$

There will be an occasion, later, to demonstrate the unbiasedness of this estimator. To understand the result at an intuitive level, one may recall that the unbiased estimator of the variance of a distribution, which is constructed from a random sample, is $\hat{\sigma}^2 = (T-1)^{-1}\sum(x_t - \hat{x})^2$. If the mean of the distribution μ were known and were used in place $\bar{x}$, then one should divide by T instead of $T-1$ to form $\hat{\sigma}^2 = T^{-1}\sum(x_t - \mu)^2$. The effect of using the datum $\bar{x}$ in place of the unknown mean μ would to reduce the measure of dispersion. To compensate, the measure is scaled by the factor $T/(T-1)$. In the context of the regression equation, where two parameters are estimated, the scale factor $T/(T-2)$ is used.

Properties of the Least-Squares Estimator

Some of the properties of the least-squares estimators that follow from the assumptions that have been made already can be revealed now. We shall consider the likelihood that these assumptions will be fulfilled in practice, as well as some consequences of their violation.

It has been assumed that the disturbance term ε is a random variable with

$$E(\varepsilon_t) = 0, \quad \text{and} \quad V(\varepsilon_t) = \sigma^2 \quad \text{for all } t. \tag{49}$$

We have avoided making statistical assumptions about x since we are unwilling to assume that its assembled values will manifest the sort of the regularities which are inherent in a statistical distribution. Therefore, the assumption that ε is independent of x cannot be expressed in terms of a joint distribution of these quantities; and, in particular, it cannot be asserted that $C(x, \varepsilon) = 0$. However, if the regressor x_t has a predetermined value that has no effect on the disturbance ε_t, then it can be stated that

$$E(x_t\varepsilon_t) = x_tE(\varepsilon_t) = 0, \quad \text{for all } t. \tag{50}$$

In place of an assumption attributing a finite variance to x, it can be assumed that

$$\lim(T \rightarrow \infty)\frac{1}{T}\sum_{t=1}^{T} x_t^2 = m_{xx} < \infty. \tag{51}$$

For the random sequence $\{x_t\varepsilon_t\}$, it can also be assumed that

$$\text{plim}(T \rightarrow \infty)\frac{1}{T}\sum_{t=1}^{T} x_t\varepsilon_t = 0. \tag{52}$$

To see the effect of these assumptions, the expression

$$y_t - \bar{y} = \beta(x_t - \bar{x}) + \varepsilon_t - \bar{\varepsilon} \tag{53}$$

may be substituted into the expression for $\hat{\beta}$ of(45). By rearranging the result, it is found that

$$\hat{\beta} = \beta + \frac{\sum(x_t - \bar{x})\varepsilon_t}{\sum(x_t - \bar{x})^2}. \tag{54}$$

The numerator of the second term on the RHS is obtained with the help of the identity

$$\begin{aligned}\sum(x_t - \bar{x})(\varepsilon_t - \bar{\varepsilon}) &= \sum(x_t\varepsilon_t - \bar{x}\varepsilon_t - x_t\bar{\varepsilon} + \bar{x}\bar{\varepsilon})\\ &= \sum(x_t - \bar{x})\varepsilon_t.\end{aligned} \tag{55}$$

From the assumption under (50), it follows that

$$E\big\{(x_t - \bar{x})\varepsilon_t\big\} = (x_t - \bar{x})E(\varepsilon_t) = 0 \quad \text{for all } t. \tag{56}$$

Therefore

$$\begin{aligned}E(\hat{\beta}) &= \beta + \frac{\sum(x_t - \bar{x})E(\varepsilon_t)}{\sum(x_t - \bar{x})^2}\\ &= \beta;\end{aligned} \tag{57}$$

and $\hat{\beta}$ is seen to be an unbiased estimator of β.

The consistency of the estimator follows, likewise, from the assumptions under (51) and (52). Thus

$$\begin{aligned}\text{plim}(\hat{\beta}) &= \beta + \frac{\text{plim}\Big\{T^{-1}\sum(x_t - \bar{x})\varepsilon_t\Big\}}{\text{plim}\Big\{T^{-1}\sum(x_t - \bar{x})^2\Big\}}\\ &= \beta;\end{aligned} \tag{58}$$

and $\hat{\beta}$ is seen to be a consistent estimator of β.

The consistency of $\hat{\beta}$ depends crucially upon the assumption that the disturbance term is independent of, or uncorrelated with, the explanatory variable or regressor x. In many econometric contexts, one should be particularly wary of this assumption. For, as we have suggested earlier, the disturbance term is liable to be compounded from the variables that have been omitted from the equation that explains y in terms of x. In a time-dependent context, these variables are liable to be correlated amongst themselves; and there may be little justification for assuming that they are not likewise correlated with x.

There are other reasons of a more subtle nature for why the assumption of the independence of ε and x may be violated. The following example illustrates one of the classical problems of econometrics.

Example. In elementary macroeconomic theory, a simple model of the economy is postulated that comprises two equations:

$$y = c + i, \tag{59}$$

$$c = \alpha + \beta y + \varepsilon. \tag{60}$$

Here, y stands for the gross product of the economy, which is also the income of consumers, i stands for investment and c stands for consumption. An additional identity $s = y - c$ or $s = i$, where s is savings, is also entailed. The disturbance term ε, which is omitted from the usual presentation in economics textbooks, is assumed to be independent of the variable i.

On substituting the consumption function of (60) into the income identity of (59) and rearranging the result, we find that

$$y = \frac{1}{1-\beta}\big(\alpha + i + \varepsilon\big), \tag{61}$$

from which

$$y_t - \bar{y} = \frac{1}{1-\beta}\big(i_t - \bar{i} + \varepsilon_t - \bar{\varepsilon}\big). \tag{62}$$

The ordinary least-squares estimator of the parameter β, which is called the marginal propensity to consume, gives rise to the following equation:

$$\hat{\beta} = \beta + \frac{\sum (y_t - \bar{y})\varepsilon_t}{\sum (y_t - \bar{y})^2}. \tag{63}$$

Equation (61), which shows that y is dependent on ε, suggests that $\hat{\beta}$ cannot be a consistent estimator of β.

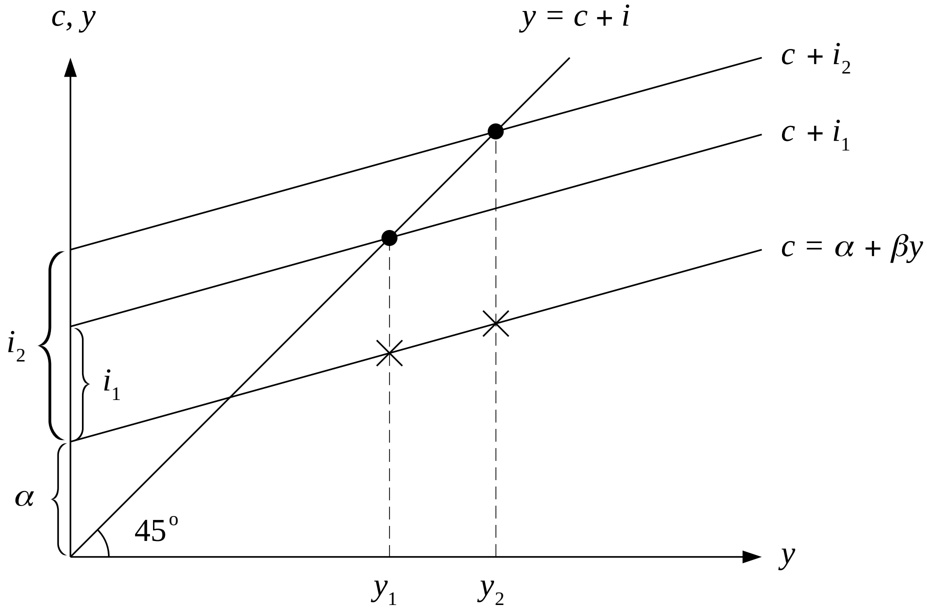

Figure 2. If the only source of variation in y is the variation in i, then the observations on y and c will delineate the consumption function.

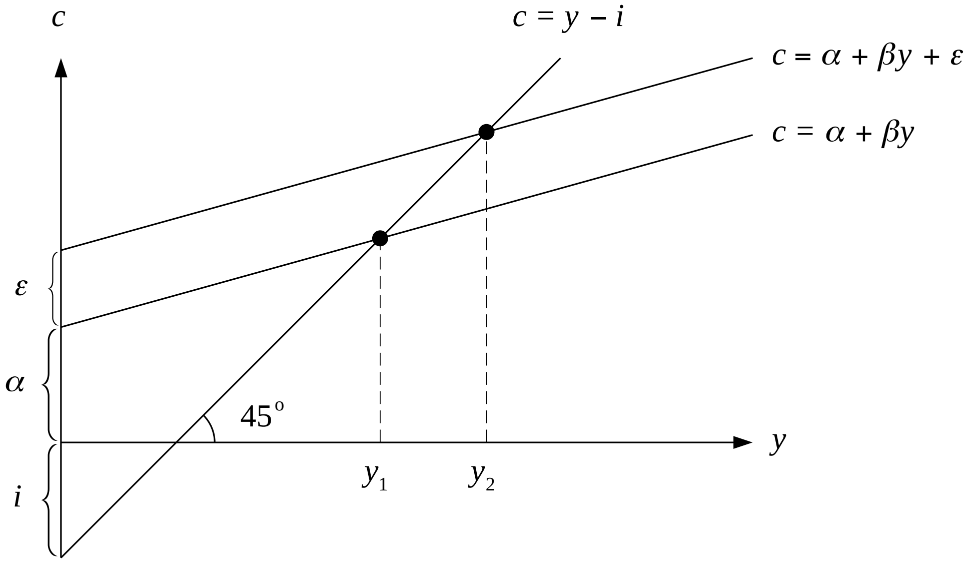

Figure 3. If the only source of variation in y are the disturbances to c, then the observations on y and c will line along a 45° line.

To determine the probability limit of the estimator, the separate probability limits of the numerator and the denominator of the term on the RHS of (63) must be determined.

The following results are available:

$$
\begin{aligned}
&\lim \frac{1}{T}\sum_{t=1}^{T}(i_t-\bar{i})^2 = m_{ii} = V(i),\\
(64)\qquad &\operatorname{plim}\frac{1}{T}\sum_{t=1}^{T}(y_t-\bar{y})^2 = \frac{m_{ii}+\sigma^2}{(1-\beta)^2} = V(y),\\
&\operatorname{plim}\frac{1}{T}\sum_{t=1}^{T}(y_t-\bar{y})\varepsilon_t = \frac{\sigma^2}{1-\beta} = C(y,\varepsilon).
\end{aligned}
$$

The results indicate that

$$
\begin{aligned}
(65)\qquad \operatorname{plim}\hat{\beta} &= \beta + \frac{\sigma^2(1-\beta)}{m_{ii}+\sigma^2}\\
&= \frac{\beta m_{ii}+\sigma^2}{m_{ii}+\sigma^2};
\end{aligned}
$$

and it can be seen that the limiting value of $\hat{\beta}$ has an upward bias which increases as the ratio σ^2/m_{ii} increases.

On the assumption that the model is valid, it is easy to understand why the parameter of the regression of c on y exceeds the value of the marginal propensity to consume. This can be achieved by considering the extreme cases.

Imagine, first, that $\sigma^2 = V(\varepsilon) = 0$. Then, the only source of variation in y and c is the variation in i. In that case, the parameter of the regression of c on y will coincide with β. This is illustrated in Figure 1. Now imagine, instead, that i is constant and that the only variations in c and y are due ε which is disturbs consumption. Then, the expected value of consumption is provided by the equation $c = y - i$ in which the coefficient associated with y is unity. Figure 2 illustrates this case. Assuming now that both $m_{ii} > 0$ and $\sigma^2 > 0$, it follows that the value of the regression parameter must lie somewhere in the interval $[\beta, 1]$.

Although it may be inappropriate for estimating the structural parameter β, the direct regression of c on y does provide the conditional expectation $E(c|y)$; and this endows it with a validity which it retains even if the Keynesian model of (59) and (60) is misspecified.

In fact, the simple Keynesian model of (59) and (60) is more an epigram than a serious scientific theory. Common sense dictates that one should give more credence to the estimate of the conditional expectation $E(c|y)$ than to a putative estimate of the marginal propensity to consume devised within the context of a doubtful model.

The Method of Maximum Likelihood

The method of maximum-likelihood constitutes a principle of estimation that may be applied to a wide variety of problems. One of the attractions of the method is that, granted the fulfilment of the assumptions on which it is based, it can be shown that the resulting estimates have optimal properties. In general, it can be shown that, at least in large samples, the variance of the resulting estimates is the least that can be achieved by any method.

The cost of using the method is the need to make the assumptions which are necessary to sustain it. It is often difficult to assess, without a great deal of further analysis, the extent to which the desirable properties of the maximum-likelihood estimators survive when these assumptions are not fulfilled. In the case of the regression model, there is considerable knowledge on this account, some of which will be presented in later chapters.

The method can be applied the regression model with independently and identically distributed disturbances that follow a normal probability law. The probability density functions of the individual disturbances $\varepsilon_t; t = 1, \ldots, T$ are given by

$$(66) \qquad N(\varepsilon_t; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right).$$

Since the ε's are assumed to be independently distributed, their joint probability density function (p.d.f.) is

$$(67) \qquad \prod_{t=1}^{T} N(\varepsilon_t; 0, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\left(\frac{-1}{2\sigma^2} \sum_{t=1}^{T} \varepsilon^2\right).$$

If the elements $x_1, \ldots, x_T$ are regarded as a given set of numbers, then it follows that the conditional p.d.f. of the sample $y_1, \ldots, y_T$ is

$$(68) \quad f(y_1, \ldots, y_T | x_1, \ldots, x_T) = (2\pi\sigma^2)^{-T/2} \exp\left\{\frac{-1}{2\sigma^2} \sum_{t=1}^{T} (y_t - \alpha - \beta x_t)\right\}.$$

The principle of maximum likelihood suggests that estimate α, β and σ^2 should be estimated by finding the values that maximise the probability measure that is attributed to the sample $y_1, \ldots, y_T$. That is to say, the events that have generated the sample are regarded as the most likely of all the events that could have occurred.

Notice that, when α, β and σ^2 are the arguments of the function f rather than its parameters, and when $y_1, \ldots, y_T$ are data values rather than random variables, the function is no longer a probability density function. For this reason, it is apt to be called a likelihood function instead. The notation by $L(\alpha, \beta, \sigma^2)$ signifies that it is a function of the variable parameters.

The log of the likelihood function, which has the same maximising values as the original function, is

$$\text{(69)} \qquad \log L = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2.$$

Given the value of σ^2, the likelihood is maximised by the values $\hat{\alpha}$ and $\hat{\beta}$ that minimise the sum of squares; and the expressions for $\hat{\alpha}$ and $\hat{\beta}$ are already available under (42) and (45) respectively.

The maximum-likelihood estimator for σ^2 can be obtained from the following first-order condition:

$$\text{(70)} \qquad \frac{\partial \log L}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2 = 0.$$

By multiplying throughout by $2\sigma^4/T$ and rearranging the result, the following estimating equation is derived:

$$\text{(71)} \qquad \sigma^2(\alpha, \beta) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)^2.$$

By putting $\hat{\alpha}$ and $\hat{\beta}$ in place, the estimator $\tilde{\sigma}^2 = \sigma^2(\hat{\alpha}, \hat{\beta}) = T^{-1}\sum e_t^2$ is derived, which has been given already under (46).

References

Galton, F., (1869), *Hereditary Genius: An Inquiry into its Laws and Consequences'* London: Macmillan.

Galton, F., (1886), Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute of Great Britain and Ireland,* 15, 246263.

Pearson, K., (1896). Mathematical contributions to the mathematical theory of evolution.III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London,* 187 , 253318.

Pearson, K., and A. Lee, (1896). Mathematical contributions to the theory of evolution. On telegony in man, &c. *Proceedings of the Royal Society of London,* 60, 273–283.

Pearson, K., and A.Lee, (1903). On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika,* 2 (4) , 357–462.