ELEMENTARY REGRESSION ANALYSIS

We shall consider three methods for estimating statistical parameters. These are the *method of moments, the method of least squares* and *the principle of maximum likelihood.*

In the case of the regression model, the three methods generate estimating equations that are identical; but the assumptions differ.

Conditional Expectations

If $y \sim f(y)$, then, in the absence of further information, the minimum-mean-square-error predictor is its expected value

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

Proof. If π is the value of a prediction, then the mean-square error is

$$(1) \qquad M = \int (y-\pi)^2 f(y) dy = E\{(y-\pi)^2\} = E(y^2) - 2\pi E(y) + \pi^2;$$

and, by calculus, it can be shown that M is minimised by taking $\pi = E(y)$.

If x is related to y, then the m.m.s.e prediction of y is the conditional expectation

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

Proof. Let $\hat{y} = E(y|x)$ and let $\pi = \pi(x)$ be any other estimator. Then,

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

In the second term, there is

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

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

$$E(xy) = \int_x \int_y xyf(x,y)\partial y \partial x = \int_x x\left\{ \int_y yf(y|x)\partial y \right\} f(x)\partial x = E(x\hat{y}).$$

When $E(xy) = E(x\hat{y})$ is rewritten as $E\{x(y-\hat{y})\} = 0$, it may be described as an orthogonality condition. This indicates that the prediction error $y-\hat{y}$ is uncorrelated with x. If it were correlated with x, then we should not be using the information of x efficiently in forming $\hat{y}$.

Linear Regression

Assume that x and y have a joint normal distribution, which implies that there is a linear regression relationship:

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

The object is to express α and β in terms of the expectations $E(x)$, $E(y)$, the variances $V(x)$, $V(y)$ and the covariance $C(x, y)$.

First, multiply (9) by $f(x)$, and integrate with respect to x to give

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

whence the equation for the intercept is

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

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

By putting (11) into $E(y|x) = \alpha + \beta x$ from (9), we find that

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

Now multiply (9) by x and $f(x)$ and integrate with respect to x to give

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

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

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

$$(13) \qquad E(xy) = \alpha E(x) + \beta E(x^2).$$

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

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

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

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

which implies that

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

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

Estimation by the Method of Moments

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ be a sample of T observations. Then, we can calculate the following empirical or sample moments:

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t, \qquad \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t,$$

$$s_x^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})^2 = \frac{1}{T}\sum_{t=1}^{T} x_t^2 - \bar{x}^2, \tag{21}$$

$$s_{xy} = \frac{1}{T}\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y}) = \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \bar{x}\bar{y}.$$

To estimate α and β, we replace the population moments in the formulae of (11) and (16) by the sample moments. Then, the estimates are

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

Convergence

We can expect the sample moments to converge to the true moments of the bivariate distribution, thereby causing the estimates of the parameters to converge likewise to the true values.

(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 is not appropriate to a stochastic sequence, such as a sequence of estimates. For, it is always possible for a_n to break the bounds of $a \pm \epsilon$ when $n > N$. The following is a more appropriate definition:

(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 $n \to \infty$ or, equivalently, $\lim P(|a_n - a| \leq \epsilon) = 1$.

With the increasing size of the sample, it becomes *virtually certain* that a_n will 'fall within an epsilon of a.' We describe a as the probability limit of a_n and we write $\text{plim}(a_n) = a$.

This definition does not presuppose that a_n has a finite variance or even a finite mean. However, if a_n does have finite moments, then we may talk of mean-square convergence:

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

We should note that

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

Thus, the mean-square error of an estimator 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 implies convergence in probability. When an estimator converges in probability to the parameter which it purports to represent, then we say that it is a consistent estimator.

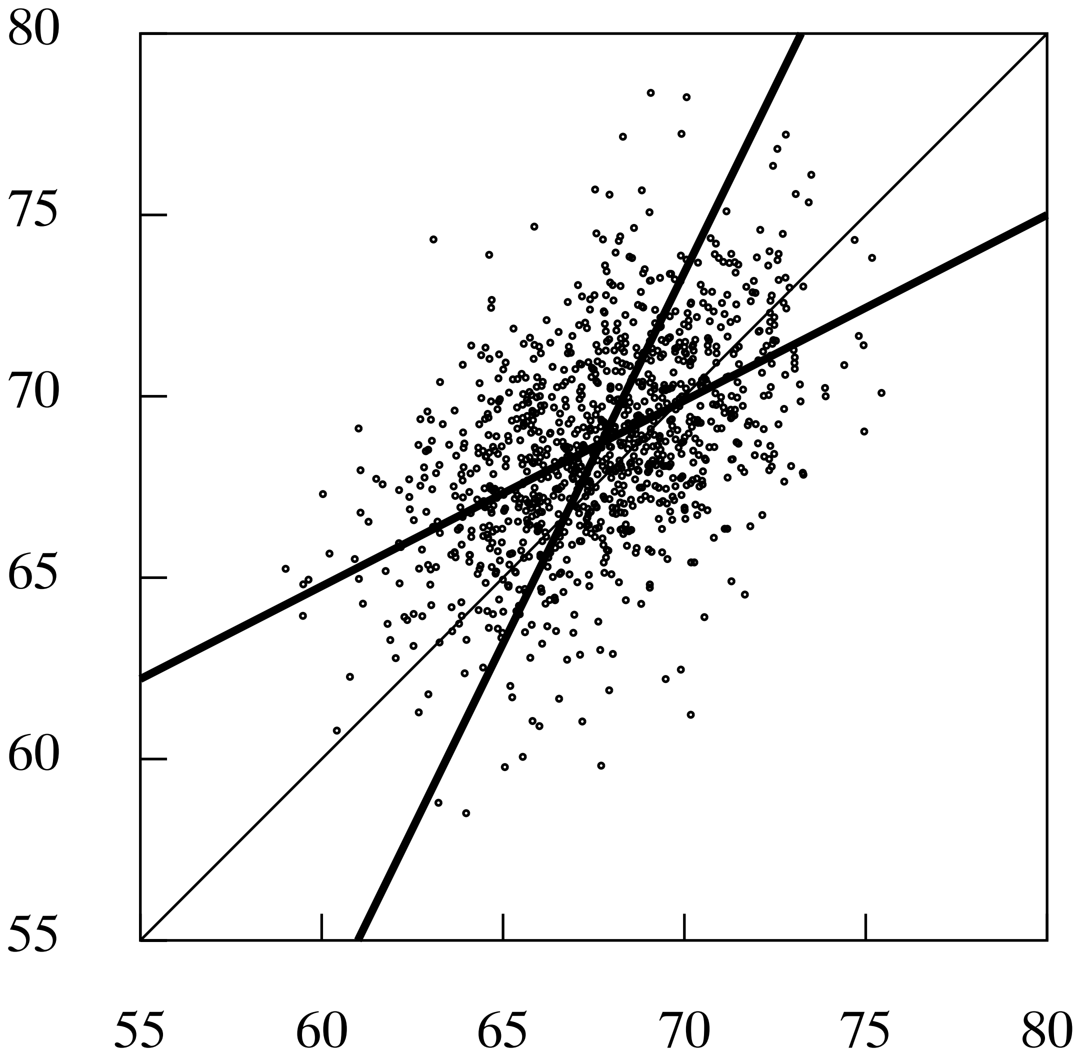

Figure 1. Pearson's data comprising 1078 measurements of the heights of fathers (the abscissae) and of their sons (the ordinates), together with the two regression lines. The correlation coefficient is 0.5013.

The Bivariate Normal Distribution

Most of the results in the theory of regression can be obtained by examining the functional form of the bivariate normal distribution. Let x and y be the two variables. Let us denote their means by $E(x) = \mu_x$, $E(y) = \mu_y$, their variances by $V(x) = \sigma_x^2, V(y) = \sigma_y^2$ and their covariance by $C(x, y) = \rho\sigma_x\sigma_y$. Here, the correlation coefficient

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

provides a measure of the relatedness of these variables.

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

(32)

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

The quadratic function can also be written as

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

Thus, we have

$$(34)\qquad f(x,y) = f(y|x)f(x),$$

where

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

and

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

with

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

Least-Squares Regression Analysis

The regression equation, $E(y|x) = \alpha + \beta x$ can be written as

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

where $\varepsilon = y - E(y|x)$ is a random variable, with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$, that is independent of x.

Given observations $(x_1, y_1), \ldots, (x_T, y_T)$, the estimates are the values that minimise the sum of squares of the distances—measured parallel to the y-axis—of the data points from an interpolated regression line:

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

Differentiating S with respect to α and setting to zero gives

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

This generates the following estimating equation for α:

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

By differentiating with respect to β and setting the result to zero, we get

$$(43) \qquad -2\sum x_t(y_t - \alpha - \beta x_t) = 0.$$

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

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

whence we get

$$(45) \qquad \hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2} = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}.$$

This is identical to the estimate under (22) derived via the method of moments. Putting $\hat{\beta}$ into the equation $\alpha(\beta) = \bar{y} - \beta\bar{x}$ of (42), gives the estimate of $\hat{\alpha}$ 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, we may apply the method of moments to the regression residuals $e_t = y_t - \hat{\alpha} - \hat{\beta}x_t$ to give

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

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

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

Properties of the Least-Squares Estimator

The disturbance term ε is assumed to be a random variable with

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

We might assume that x is a random variable uncorrected with ε such that that $C(x, \varepsilon) = 0$. However, if we are prepared to regard the x_t as predetermined values which have no effect on the ε_t, then we can say that

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

In place of an assumption attributing a finite variance to x, we may assert that

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

For the random sequence $\{x_t \varepsilon_t\}$, we assert that

$$(52) \qquad \operatorname{plim}(T \to \infty) \frac{1}{T} \sum_{t=1}^{T} x_t \varepsilon_t = 0.$$

To see the effect of these assumptions, let us substitute the expression

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

in the expression for $\hat{\beta}$ found under (45). By rearranging the result, we have

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

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,

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

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

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

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.

Example. A simple model of the economy is postulated that comprises two equations in income y, consumption c and investment i:

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

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

Also, $s = y - c$ or $s = i$, where s is savings. The disturbance ε, is assumed to be independent of investment i. Substituting (60) into (59) gives

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

from which

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

The estimator of the parameter β, marginal propensity to consume is

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

Since y is dependent on ε, according to (61), $\hat{\beta}$ cannot be a consistent estimator of β.

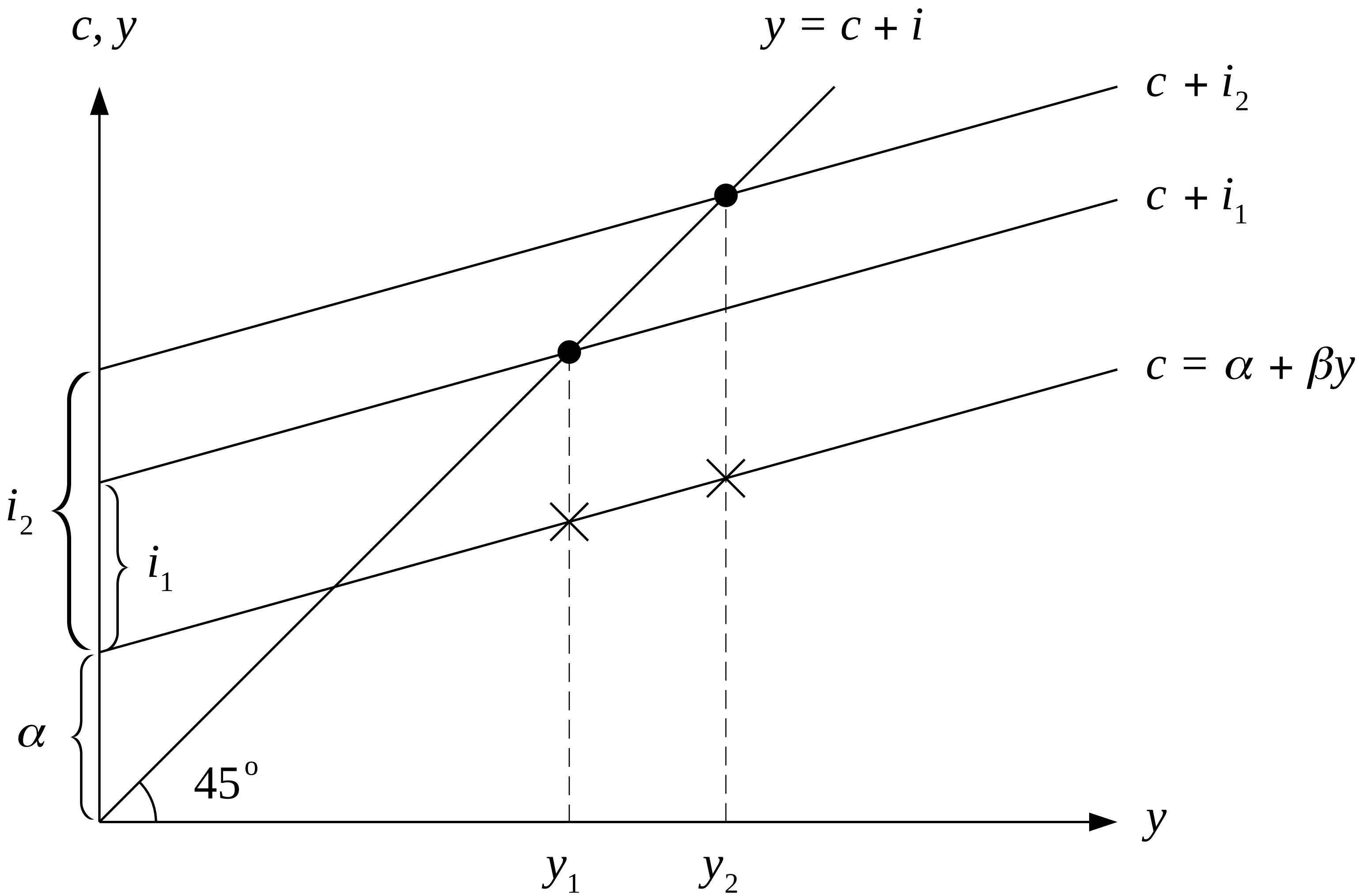

Figure 2. If the only source of variation in y is the variation in i, then the observations on y and c will delineate the consumption function.

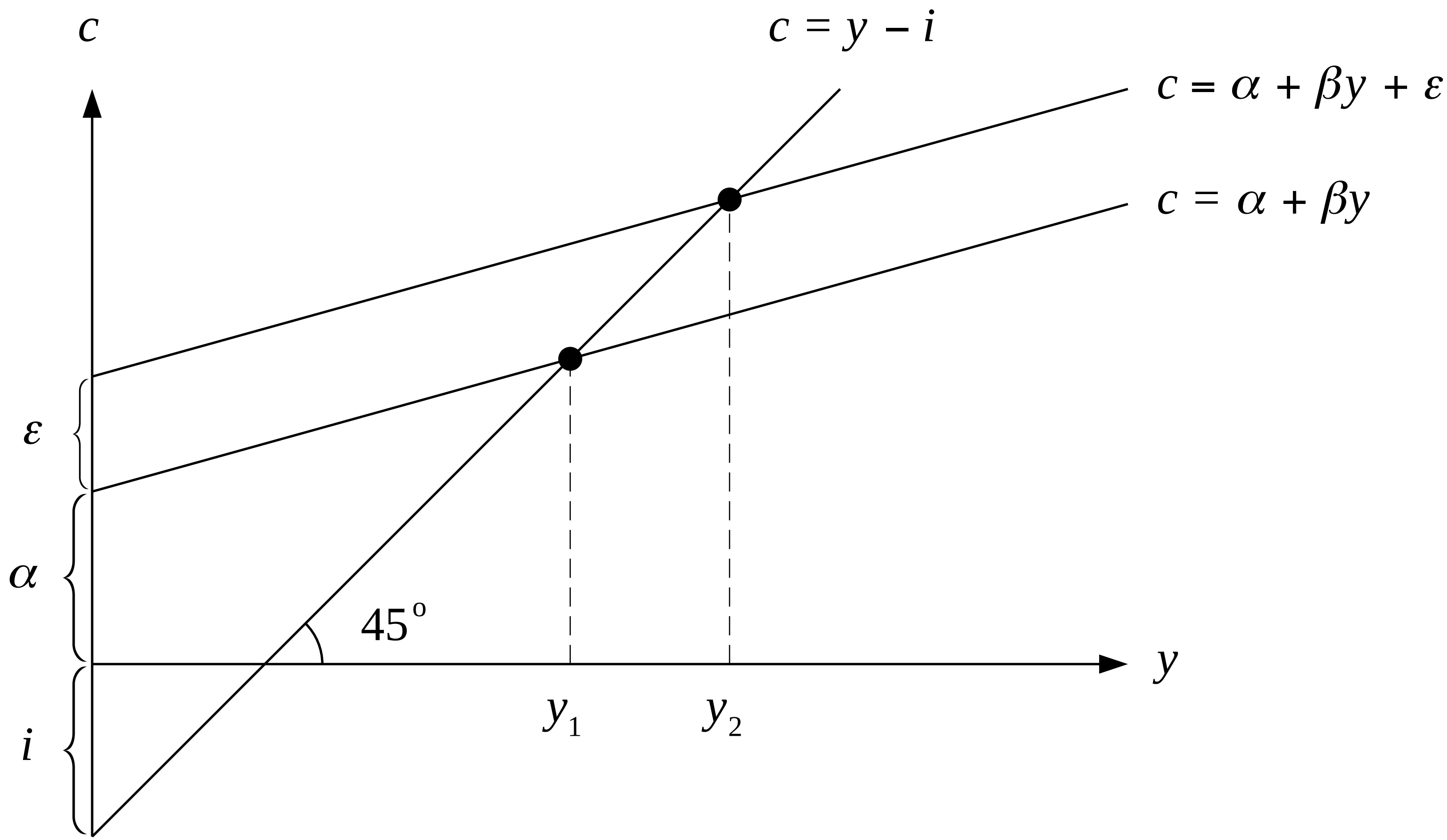

Figure 3. If the only source of variation in y are the disturbances to c, then the observations on y and c will line along a 45° line.

To determine the probability limit of the estimator, we must assess the separate probability limits of the numerator and the denominator of the term on the RHS of (63). The following results are available:

$$\lim \frac{1}{T}\sum_{t=1}^{T}(i_t - \bar{i})^2 = m_{ii} = V(i),$$

$$\text{plim}\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2 = \frac{m_{ii} + \sigma^2}{(1-\beta)^2} = V(y), \tag{64}$$

$$\text{plim}\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})\varepsilon_t = \frac{\sigma^2}{1-\beta} = C(y, \varepsilon).$$

The results indicate that

$$\text{plim}\ \hat{\beta} = \beta + \frac{\sigma^2(1-\beta)}{m_{ii} + \sigma^2} = \frac{\beta m_{ii} + \sigma^2}{m_{ii} + \sigma^2}. \tag{65}$$

The Method of Maximum Likelihood

The disturbance $\varepsilon_t; t = 1, \ldots, T$ in the regression model are assumed to be independently and identically distributed with a normal density:

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

Since they 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 we regard the elements $x_1, \ldots, x_T$ 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)

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

The maximum likelihood estimates α, β and σ^2 are the values that maximise the probability measure that is attributed to the sample $y_1, \ldots, y_T$. The log likelihood function, which is maximised by these values, is

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

Given the value of σ^2, this is maximised by the values $\hat{\alpha}$ and $\hat{\beta}$ under (42) and (45) respectively, which minimise the error sum of squares.

The estimate of σ^2 is from the following first-order condition:

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

Multiplying throughout by $2\sigma^4/T$ and rearranging the result, gives

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