ECONOMETICS FOR THE UNINITIATED

by D.S.G. POLLOCK
University of Leicester

This lecture can be supplemented by the texts that are available at the following website, where each of the topics is pursued in greater detail:

`http://www.le.ac.uk/users/dsgp1/`

Reference should be made to the following items:

2. Introductory Econometrics,
3. Intermediate Econometrics,
7. EC 3062 Econometric Theory.

1. An Econometric Regression Equation

Consider the expenditure on food and clothing of a group of individual households observed over a given period. We might postulate that

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

where

y_i is expenditure of the ith family,
x_i is its income and
ε_i is a random variable.

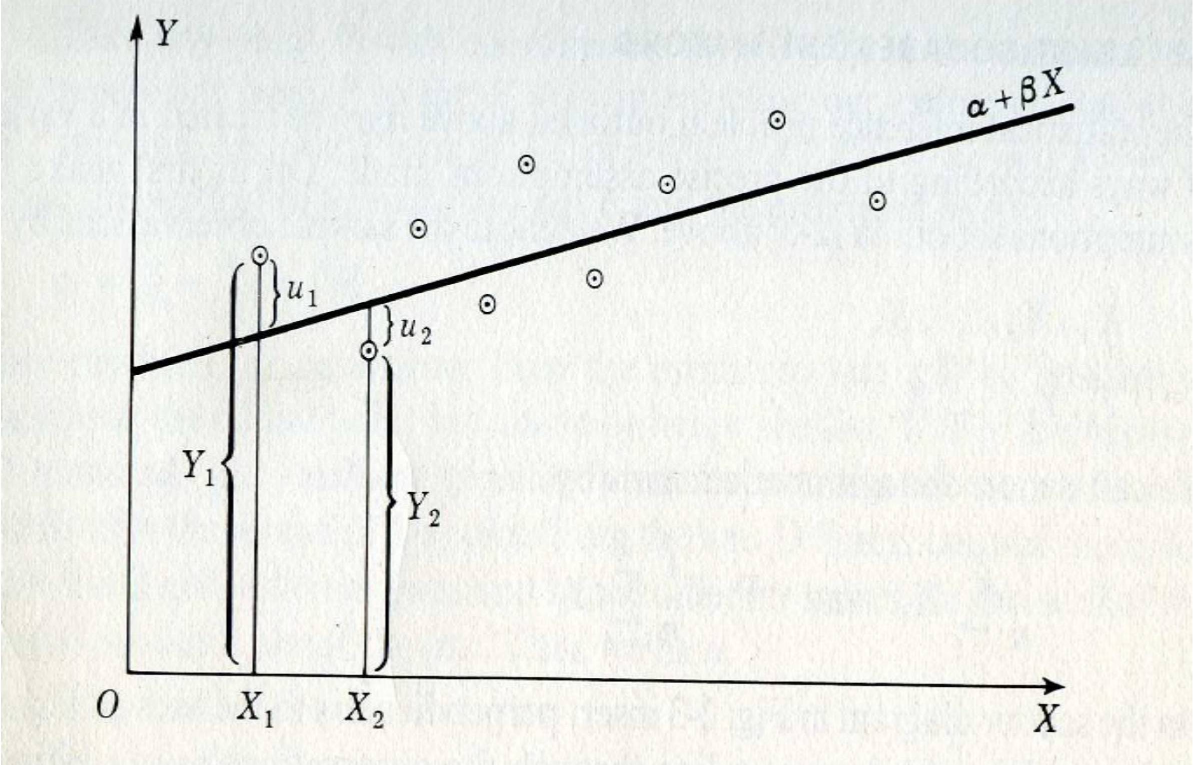
The random variable ε_i has its own tenuous regularities which can be summarised by the parameters of a statistical distribution.

It might be assumed that ε_i is independently and identically distributed for all i, with its expectation of $E(\varepsilon_i)$ and its variance of $V(\varepsilon_i)$ given by

$$E(\varepsilon_i) = 0 \quad \text{and} \quad V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2.$$

Under an alternative assumption, which might be more realistic, there would be

$$V(\varepsilon_i) = x_i\sigma^2.$$



A simple linear regression equation with a slope parameter β and an intercept parameter α

2. Bivariate Distributions

The joint density function of x and y is

$$f(x,y)=f(x|y)f(y)=f(y|x)f(x),$$

where

$$f(x)=\int_y f(x,y)dy \qquad \text{and} \qquad f(y)=\int_x f(x,y)dx$$

are the marginal distributions of x and y respectively, and where

$$f(x|y)=\frac{f(y,x)}{f(y)} \qquad \text{and} \qquad f(y|x)=\frac{f(y,x)}{f(x)}$$

are the conditional distributions of x given y and of y given x.

The unconditional expectations of x and y are

$$E(x)=\int_x xf(x)dx \quad \text{and} \quad E(y)=\int_y yf(y)dy.$$

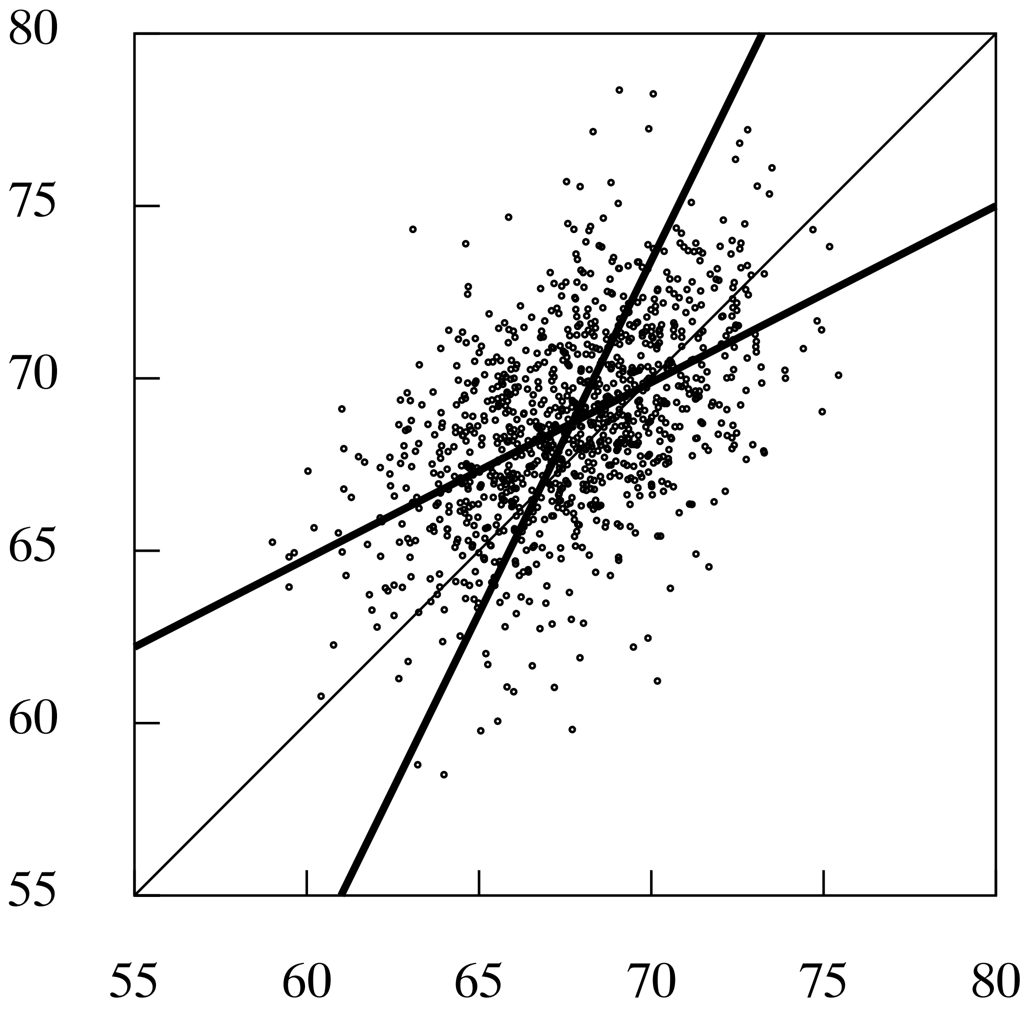
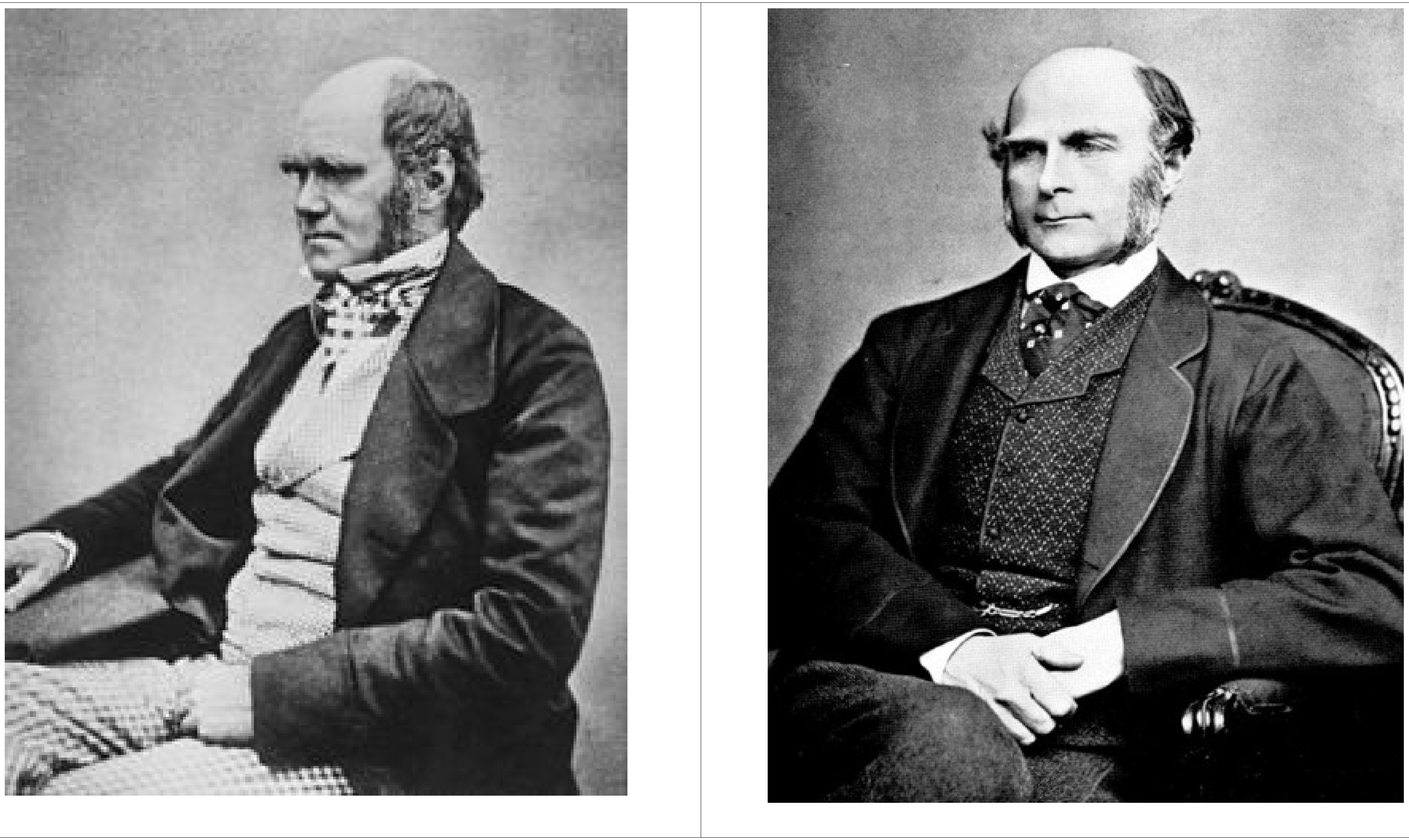
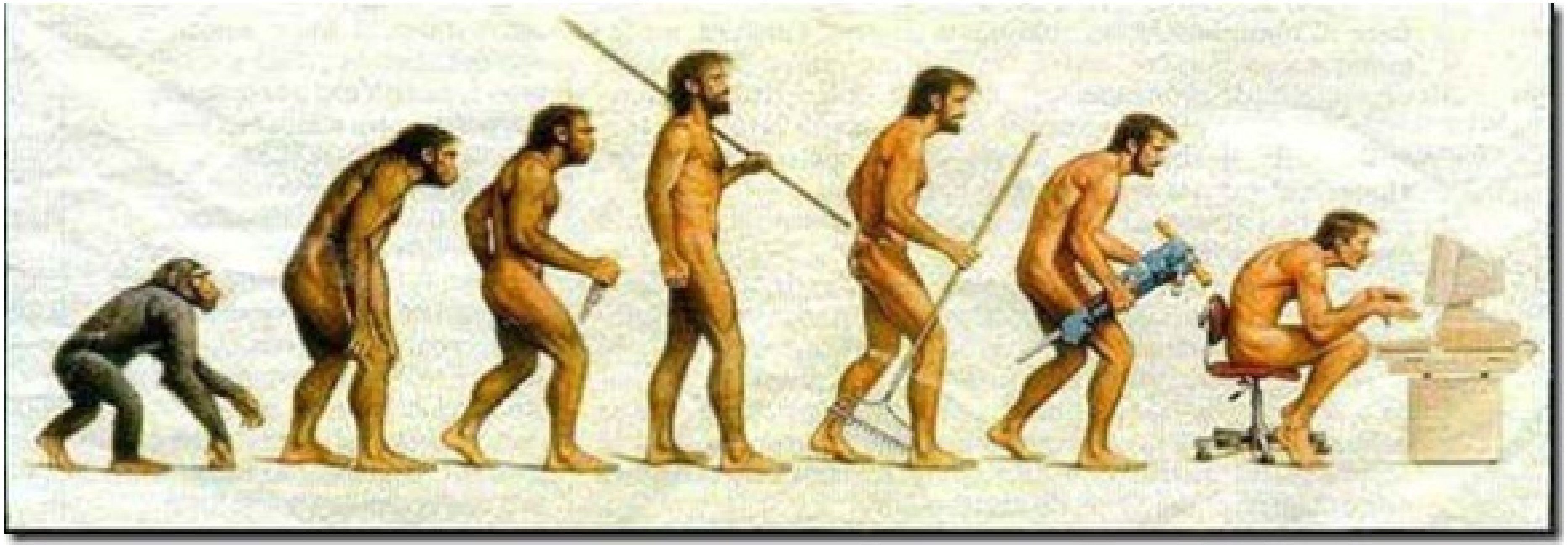


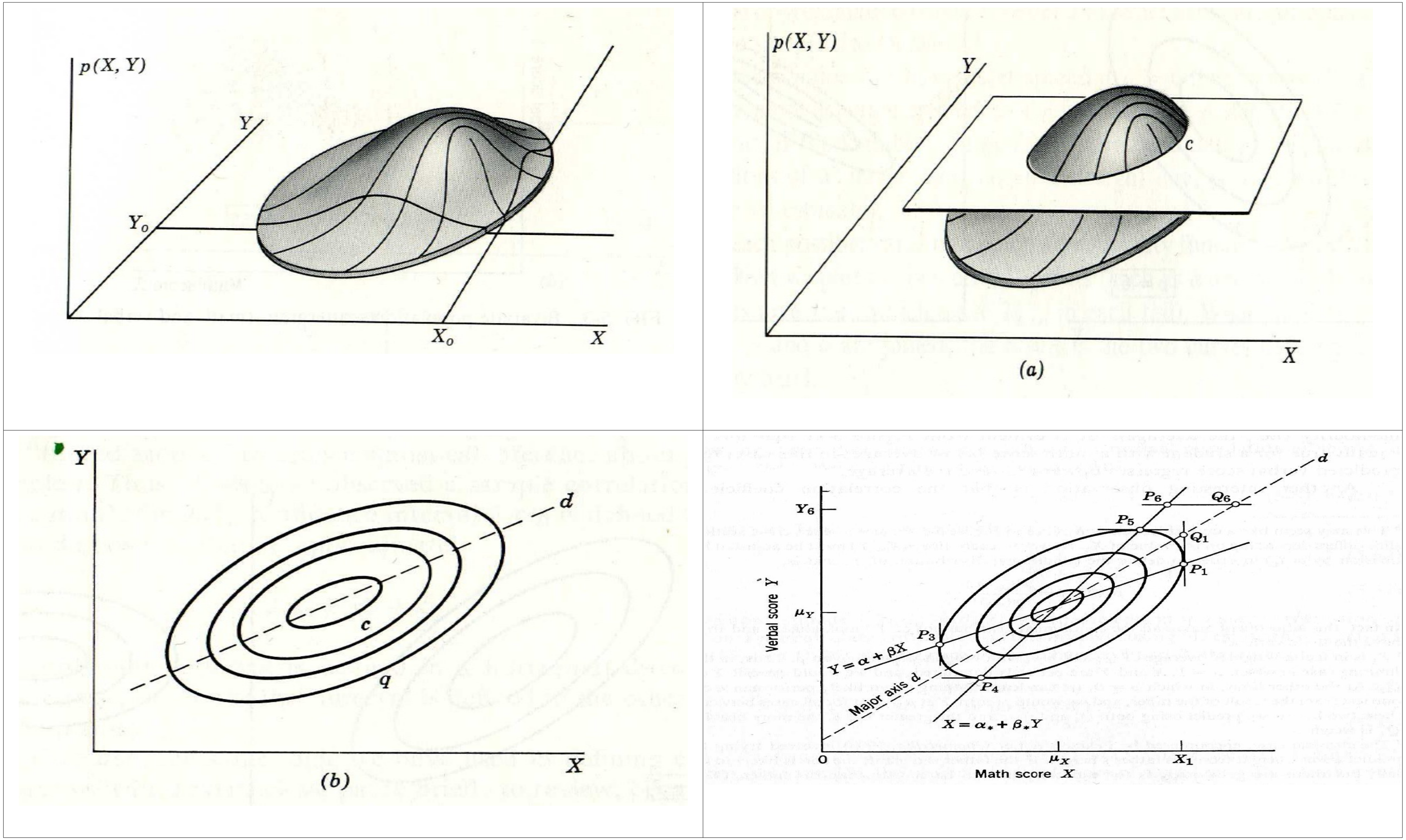
Figure 1. Pearson's data comprising 1078 measurements of on the heights of father (the abscissae) and of their sons (the ordinates), together with the two regression lines. The correlation coefficient is 0.5013.



Charles Darwin (1809--1882) and Francis Galton (1811--1911)



The Ascent of Man in a Modern Perspective



The Bivariate Normal Gaussian Distribution

3. Conditional Expectations and Regression

The conditional expectation of y given x is

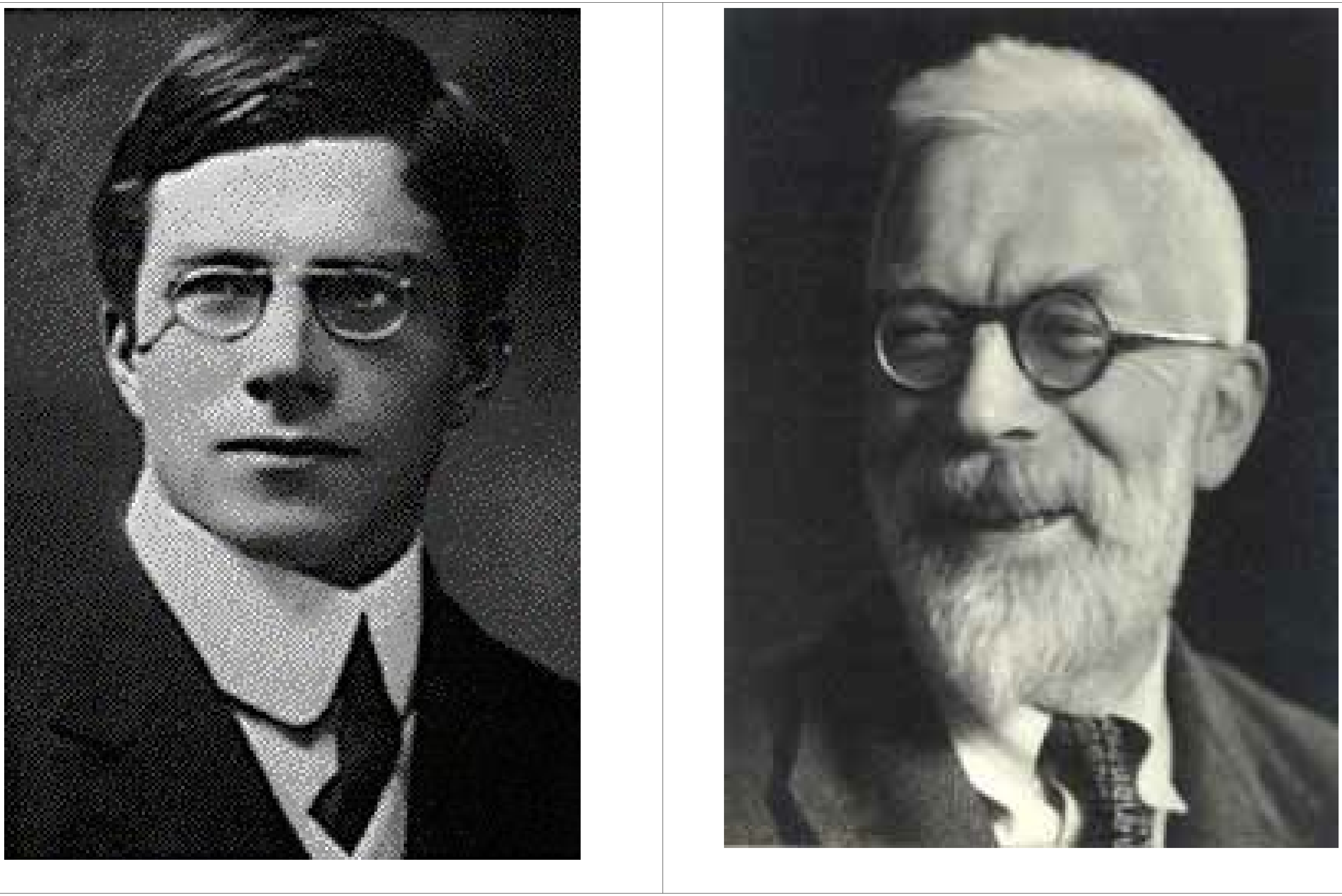
$$E(y|x) = \int_y yf(y|x)dy = \int_y y\frac{f(y,x)}{f(x)}dy.$$

In the case of the bivariate normal distribution, this constitutes a linear regression:

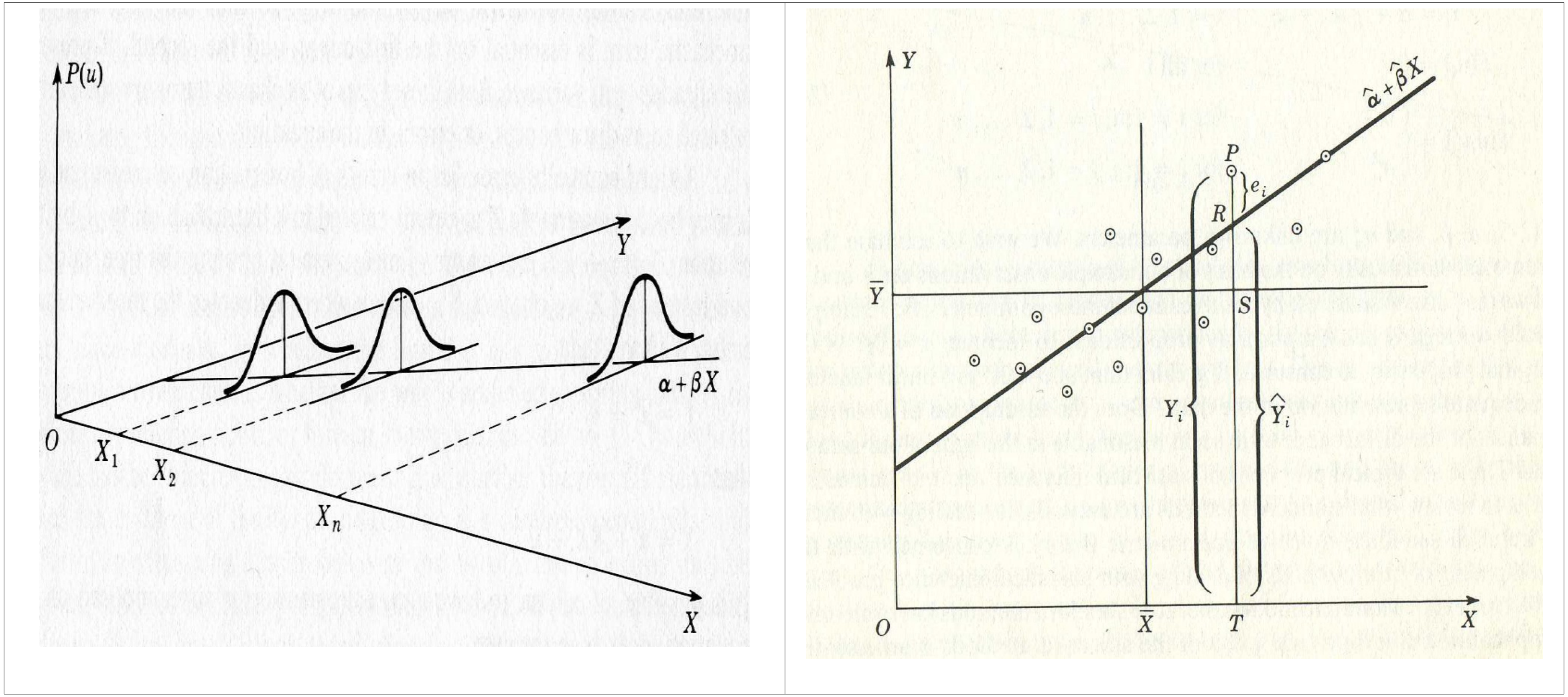
$$\begin{aligned} E(y|x) &= E(y) + \beta\{x - E(x)\} \\ &= \{E(y) - \beta E(x)\} + \beta x \\ &= \alpha + \beta x. \end{aligned}$$

Here, $\alpha = E(y) - \beta E(x)$; and it can be shown that

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



Images of R.A. Fisher (1890--1962)



The conditional expectations and the least-squares estimation of the regression line

4. Ordinary Least-Squares Regression Estimates

It is customary, in econometric texts, to derive estimates of the regression parameters α and β by minimising the sum of squares of the residual deviations:

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

By solving the conditions obtained by differentiating S with respect to α and β and by setting the results to zero, it is found that

$$\alpha(\beta) = \bar{y} - \beta\bar{x},$$

and that

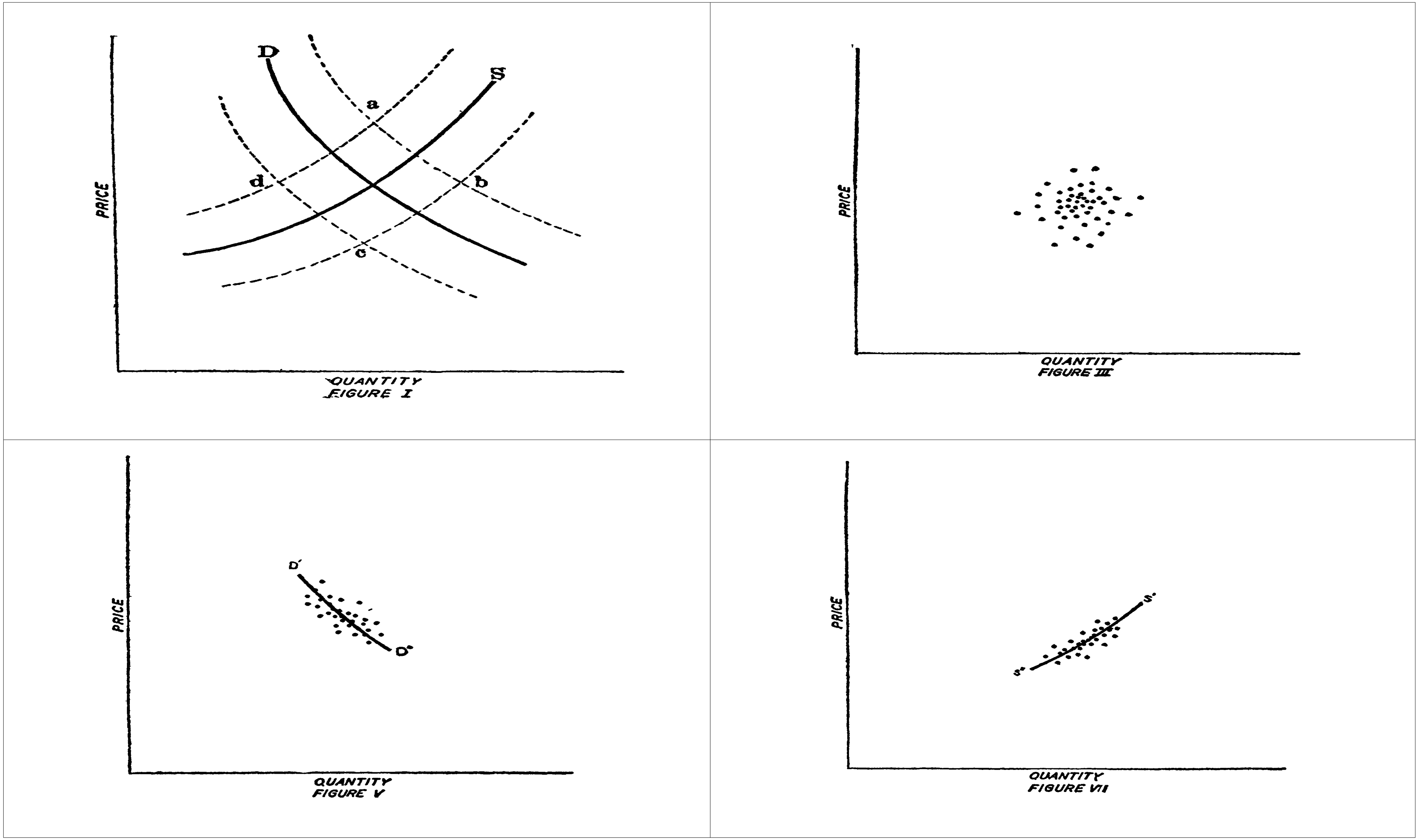
$$\hat{\beta} = \frac{\sum x_t y_t - T\bar{x}\bar{y}}{\sum x_t^2 - T\bar{x}^2} = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sum (x_t - \bar{x})^2}.$$

These estimates can also be obtained by replacing the theoretical moments $E(x)$, $E(y)$, $V(x)$ and $C(x, y)$, within the expressions for α and β, by their empirical counterparts

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t, \quad \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t, \quad s_x^2 = \frac{1}{T}\sum_{t=1}^{T} (x_t - \bar{x})^2, \quad s_{xy} = \frac{1}{T}\sum_{t=1}^{T} (x_t - \bar{x})(y_t - \bar{y}).$$

Marshallian Scissors

We might as reasonably dispute whether it is the upper or the under blade of a pair of scissors that cuts a piece of paper, as whether value is governed by utility or cost of production.



Diagrams of supply and demand from the paper of E.J. Working *What do Statistical Demand Curves Show?*

5. Simultaneous Equations and the Problem of Identification

Consider the following system:

$$y_1 = y_2\gamma_{21} + \varepsilon_1 : \qquad \text{The Demand Equation,}$$

$$y_2 = y_1\gamma_{12} + x\beta + \varepsilon_2 : \qquad \text{The Supply Equation,}$$

where

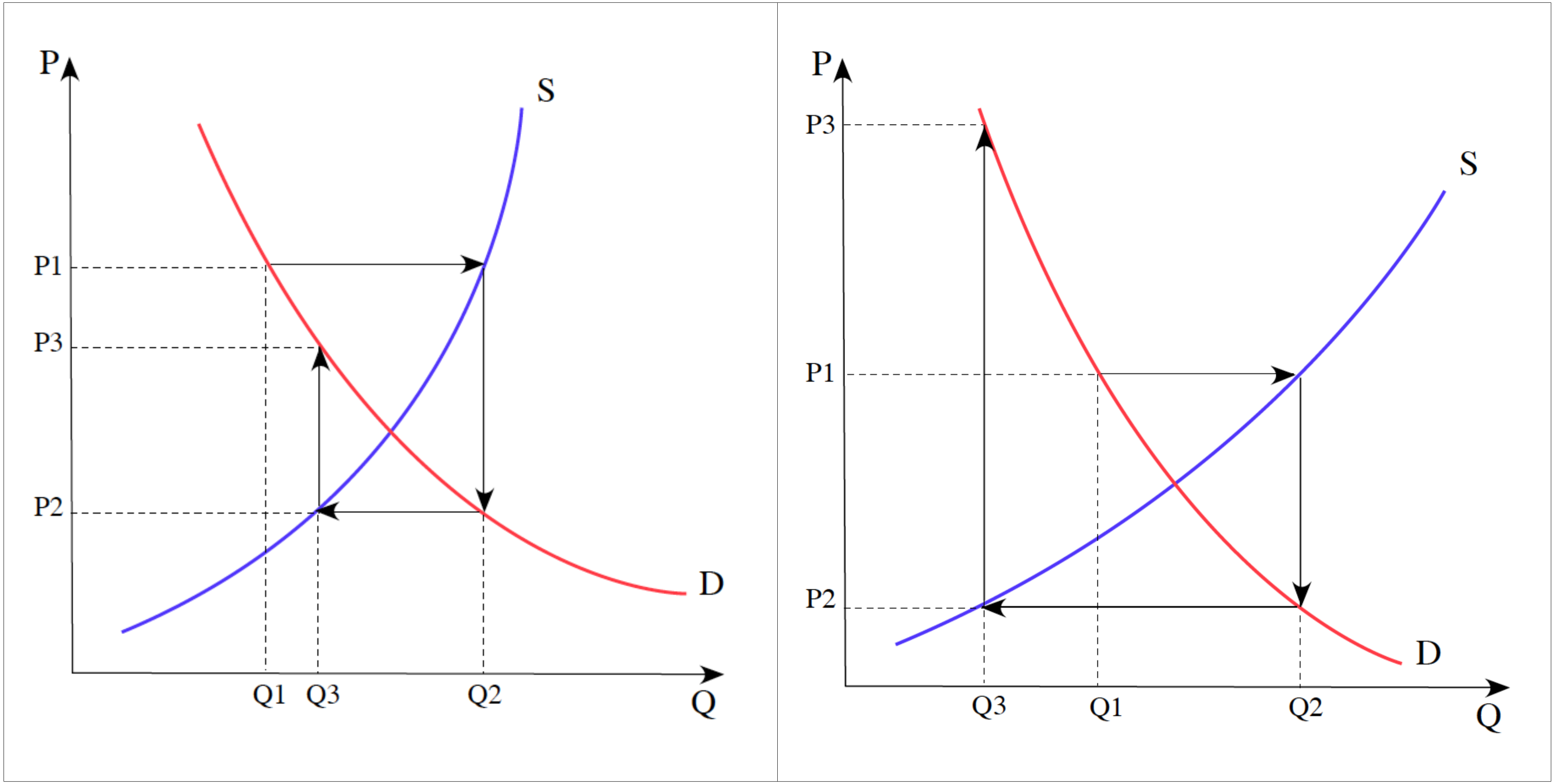
y_1 represents the quantity of popcorn consumed and produced,

y_2 represents the price of popcorn and

x represents the cost of maize.

If the exogenous variable x has a reasonable degree of variability relative to ε_1 and ε_2, and if it is statistically independent of the latter, then it will serve to shift the supply curve in such a way as to reveal the profile of the demand curve.

The supply curve itself will be identifiable only if another exogenous variable enters the demand equation.



The cobweb model, which explains how the prices and quantities in agricultural markets may show repeated annual fluctuations. The geometry of the diagrams determine whether the amplitude of the fluctuations will be increasing or diminishing

6. Recursive and Dynamic Models

The existence of a temporal ordering amongst the equations can greatly alleviate the problems of identification and estimation. Consider the following equation describing the market for a crop:

$$q_t = \beta p_{t-1} + \varepsilon_{qt} : \qquad \text{The Demand Equation,}$$

$$p_t = \gamma q_t + \varepsilon_{pt} : \qquad \text{The Supply Equation,}$$

The supply is determined in reference to last years price p_{t-1}, whereas the demand is determined by this year's quantity q_t.

Substituting for q_t in the supply equation gives

$$p_t = \gamma\beta p_{t-1} + \{\gamma\varepsilon_{qt} + \varepsilon_{pt}\}.$$

The equation will be dynamically stable, with bounded oscillations, if $|\gamma\beta| < 1$. It will be unstable if $|\gamma\beta| \geq 1$.

7. Distributed Lags

The effect of the changes in the signal variable x upon the dependent variable y may be distributed over time. The following equation might capture such an effect:

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_k x_{t-k} + \varepsilon_t.$$

The coefficients of the equation describe the so-called impulse response function

$$\{r_t\} = \big\{ \ldots, 0, \beta_0, \beta_1, \ldots, \beta_k, 0, \ldots \big\}.$$

Disregarding the effect of the disturbance term ε, this would be the response of the equation to a signal sequence of the form

$$\{x_t\} = \big\{ \ldots, 0, 1, 0, \ldots, 0, 0, \ldots \},$$

which comprises a single unit impulse preceded and followed by zero values.

A problem with this formulation is that it is wasteful in its use of parameters, which will be difficult to determine with accuracy if the data sequence is of a limited variability and if it has a high degree of inertia.

8. Geometric Lags

The geometric-lag model overcomes the problem of excessive parametrisation:

$$y_t = \beta\{x_t + \phi x_{t-1} + \cdots + \phi^{t-1} x_0\} + \phi^t \theta + \varepsilon_t.$$

Its impulse response is a geometrically declining sequence

$$\{r_t\} = \{\ldots, 0, \beta, \beta\phi, \beta\phi^2, \ldots, \beta\phi^k, \ldots\}.$$

Although the equation is not amenable to ordinary least-squares regression, it is reasonably straightforward to estimate its parameters.

A simpler way of creating a geometric lag scheme is to include a lagged value of the dependent variable on the RHS of the equation to give

$$y_t = \phi y_{t-1} + \beta x_t + \varepsilon_t.$$

By repeated substitution, if can be shown that

$$y_t = \beta \sum_{i=0}^{\infty} \phi^i x_{t-i} + \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}.$$

9. Problems with Trended Variables

Granger and Newbold showed that variables that have a high degree of inertia can be highly correlated, even when they have been generated independently. Thus, two independent processes

$$x_t = \beta x_{t-1} + \varepsilon_t \quad \text{and} \quad y_t = \gamma y_{t-1} + \eta_t$$

with γ and β close to unity will generate data with a high degree of correlation that diminishes only gradually as the sample size increases. This creates a danger of spurious regressions.

Equally, if the variable y shows a high degree of inertia, or if it is trended, then a model of the form

$$y_t = \phi y_{t-1} + \beta x_t + \varepsilon_t$$

is liable to fit the data deceptively well, since y_{t-1} will be close to y_t. Such phenomena have given rise to false claims of success in explaining economic processes.

The theory of regression requires the variables to be stationary, which implies an absence of trend. It is often found that, when the variables are detrended, e.g. by replacing them by their differences $\nabla y_t = y_t - y_{t-1}$ and $\nabla x_t = x_t - x_{t-1}$, the degree of their correlation is radically reduced.

10. Cointegrated Variables

If two trended variables maintain proportionality in the log run and if they can be reduced to stationarity by taking their differences, then it may be possible to describe their relationship in a manner that conforms to the theory of linear regression and which allows the relationship to be subject, in part, to the ordinary methods of statistical inference. In that case, the variables are said to be co-integrated.

Consider the following second-order autoregressive distributed-lag model:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t.$$

Imagine that whereas x_t is trended its difference $\nabla x_t = x_t - x_{t-1}$ is stationary, which implies that the difference $\nabla y_t = y_t - y_{t-1}$ is also stationary. Then, the equation can be recast in the following error-correction form

$$\nabla y_t = \lambda \{\gamma x_{t-1} - y_{t-1}\} + \rho \nabla y_{t-1} + \delta \nabla x_t + \varepsilon_t,$$

where

$$\lambda = 1 - \phi_1 - \phi_2, \quad \gamma = \frac{\beta_0 + \beta_1}{1 - \phi_1 - \phi_2}, \quad \rho = \phi_2, \quad \text{and} \quad \delta = \beta_1.$$

11. The Error-Correction Equation

In the reformulated error-correction equation, γ represents the coefficient of proportionality that defines the long-run relationship between y—which might denote aggregate consumption—and x—which might be aggregate disposable income.

The term $\gamma x_{t-1} - y_{t-1}$ is the disequilibrium error at time t. The error will tend to be eliminated via the adjustment ∇y_t. If warranted consumption γx_{t-1} has exceeded actual consumption y_{t-1}, then a positive impetus will be imparted to ∇y_t and, conversely, if actual consumption has exceeded the warranted value, the adjustment will be negative.

The parameter $\lambda \in [0, 1)$ governs the speed of the adjustments—the closer is its value to unity, the more rapid are the adjustments.

The original autoregressive distributed-lag model, which is in levels, may be estimated by ordinary least-squares regression. When the model has been recast into its error correction form, the parameters λ, ρ and δ are amenable to the usual tests of statistical significance, albeit that these are valid only for large samples.

The parameter γ, which governs the long-term proportionality, has a special statistical distribution, which reflects the fact that it is determined to a higher degree of accuracy than the other parameters. Its estimate is said to be super-consistent.