Chapter 1

The Linear Regression Model

1.1 The Model

Let us begin by defining some symbols and writing down some notation, without any preconceptions, initially, as to what the context of these relationships might be. Let $x_{1t}, \ldots, x_{kt}, y_t$ for $t = 1, \ldots, n$ denote a sample of n observations on $k+1$ variables, and consider the linear equation

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + u_t \tag{1.1.1}$$

where $\beta_1, \ldots, \beta_k$ are fixed but unknown parameters. The variable u_t is defined by the equation, being that function of the observables that balances the equation for observation t. If the parameters are unknown, then u_t is likewise unknown. The variable y_t is called the *regressand*, and the x_{it} variables for $i = 1, \ldots, k$ are the *regressors.* Usually one of the regressors is fixed at unity, say $x_{kt} = 1$, all t, and its coefficient β_k is called the *intercept* of the equation.

The observations are typically either a *time series*, relating to successive time periods in the study of economic aggregates, or a *cross-section* of individual economic units – households, firms, industries, countries and so forth, observed at a point of time.[1] The intercept apart, the variables in the equation, the y_t and usually the x_{it} too (subject to the considerations discussed in §1.3) are assumed to be *random variables.* This means that the observations are supposed to be generated by a random experiment, in advance of which their values are unknown. The notion of 'experiment' here is rather loose, and might refer merely to the act of collecting the sample. The relevant probability theory is summarized in Appendix B, but for present purposes it will do no harm to treat the concept of randomness in a purely intuitive fashion. It aims to capture the idea that the experiment

[1] The case of *panel data* in which the observations have both dimensions, e.g. the same sample of households observed in successive years, is treated only briefly in this book; see §2.3.1.

might in principle be performed repeatedly, in each case throwing up a new sample whose particular values are not predictable in advance. All we need to assume is that the notion of an *expected value* is well-defined, and there exist constants $E(y_t)$, $E(x_{it})$ and $E(u_t)$ representing central tendencies of the distributions of the indicated variables.

To provide a convenient shorthand form of the equation, define the vectors

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad (k \times 1) \qquad \boldsymbol{x}_t = \begin{bmatrix} x_{1t} \\ \vdots \\ x_{kt} \end{bmatrix} \quad (k \times 1)$$

and then write $y_t = \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$, or equivalently $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t$. The usual practice is to write down all the observations in one vector equation. Letting

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad (n \times 1) \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} \quad (n \times k) \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \quad (n \times 1)$$

the complete sample of equations is represented by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{1.1.2}$$

In the context of econometrics, equation (1.1.1) is usually thought of as a model of economic behaviour. The variable y_t typically represents the response of economic agents to a collection of 'stimulus' variables $\boldsymbol{x}_t$. The equation 'explains' y_t as a function of $\boldsymbol{x}_t$. Such a model is typically suggested by economic theory when agents are assumed to formulate a plan for y_t based on $\boldsymbol{x}_t$, for example, a plan for consumption expenditure given income received and prices observed. The notion of a systematic rule of behaviour is embodied in the assumption that the coefficients $\beta_1, \ldots, \beta_k$ are constants. u_t is called the *disturbance* term, or the *error* term and, in the context of the behavioural model, represents the deviations from the plan. We assume $E(u_t) = 0$, although note that unless the intercept β_k is constrained in some way, this assumption is trivially valid. By estimating $\boldsymbol{\beta}$ and setting $u_t = 0$ the model can be used to predict y_t, given predictions or observations of $\boldsymbol{x}_t$.

1.2 The Least Squares Estimator

1.2.1 Derivation of the Estimator

An *estimator* is a rule (for example, a formula) for computing an *estimate* (a number) from sample data. The method of least squares is the standard technique for extracting an estimate of $\boldsymbol{\beta}$ from a sample of observations. Consider

$$\boldsymbol{e}(\boldsymbol{b}) = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} \quad (n \times 1) \tag{1.2.1}$$

a vector of functions with k-dimensional vector argument $\boldsymbol{b}$. Choosing a vector $\boldsymbol{b}$ to represent the unknown $\boldsymbol{\beta}$, $\boldsymbol{X}\boldsymbol{b}$ may be thought of as a predictor of $\boldsymbol{y}$, and then

$\boldsymbol{e}$ is the corresponding prediction error. The sum of the squared prediction errors is

$$S(\boldsymbol{b}) = \boldsymbol{e}(\boldsymbol{b})'\boldsymbol{e}(\boldsymbol{b}). \tag{1.2.2}$$

If the criterion of a good estimator is one that yields a good predictor of $\boldsymbol{y}$ given $\boldsymbol{X}$, a natural choice is $\hat{\boldsymbol{\beta}}$ such that $S(\hat{\boldsymbol{\beta}}) \leq S(\boldsymbol{b})$ for all k-vectors $\boldsymbol{b}$, denoted more formally by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} S(\boldsymbol{b}) \tag{1.2.3}$$

where 'arg min' is shorthand for 'the argument that minimizes' the function in question. This is called the *ordinary least squares* (OLS) estimator. The corresponding estimator of $\boldsymbol{u}$ is denoted by $\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, the *least squares residuals*, such that $S(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}$.

As is well known, $\hat{\boldsymbol{\beta}}$ is obtained from the data by a simple mathematical formula. One way to derive this formula is from the following result:

Theorem 1.2.1 Necessary and sufficient conditions for a vector $\hat{\boldsymbol{\beta}}$ to be the unique minimizer of S are

(a) $\operatorname{rank}(\boldsymbol{X}) = k$

(b) $\boldsymbol{X}'\hat{\boldsymbol{u}} = \boldsymbol{0}$.

Proof Substitute $\hat{\boldsymbol{u}} + \boldsymbol{X}\hat{\boldsymbol{\beta}}$ for $\boldsymbol{y}$, so as to write

$$\begin{aligned} S(\boldsymbol{b}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \\ &= (\hat{\boldsymbol{u}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}))'(\hat{\boldsymbol{u}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b})) \\ &= S(\hat{\boldsymbol{\beta}}) + 2(\hat{\boldsymbol{\beta}} - \boldsymbol{b})'\boldsymbol{X}'\hat{\boldsymbol{u}} + (\hat{\boldsymbol{\beta}} - \boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b}). \end{aligned} \tag{1.2.4}$$

$\boldsymbol{X}'\boldsymbol{X}$ is positive definite if and only if condition (a) holds, by Lemma A.7.1, and hence $S(\boldsymbol{b})$ is minimized uniquely at the point $\hat{\boldsymbol{\beta}} = \boldsymbol{b}$ if and only if both the conditions hold, by Lemma A.7.3. ∎

Let us see what this result implies. According to the orthogonality condition (b), the sample covariances of the residuals and each regressor are identically zero, noting that since one column of $\boldsymbol{X}$ is the column of ones, the residuals also sum to 0 by construction. If this were not so, some linear function of $\boldsymbol{X}$ could be used to predict $\hat{\boldsymbol{u}}$. The least squares predictor of $\boldsymbol{y}$,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \hat{\boldsymbol{u}} \tag{1.2.5}$$

could be improved by adding this predictor to it. $\hat{\boldsymbol{\beta}}$ is the linear estimator which cannot be so improved, for all such corrections have been made. It can easily be verified that

$$\hat{\boldsymbol{u}}'\hat{\boldsymbol{y}} = 0. \tag{1.2.6}$$

It is straightforward to obtain the formula for $\hat{\boldsymbol{\beta}}$ from condition (b) of Theorem 1.2.1, which may be written as

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{1.2.7}$$

These are the so-called *normal equations* of least squares. Noting that $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular by Lemma A.7.1, they have the unique solution

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{1.2.8}$$

This completes the derivation of the OLS estimator. Notice that calculus has not been used. Solving the first-order conditions for the minimization of S is an alternative and more common way to obtain the formula, and is an easy exercise using the results of §A.9.

1.2.2 Goodness of Fit

Condition (1.2.6) implies the well-known *sum of squares decomposition*,

$$\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}. \tag{1.2.9}$$

The square of the sample correlation coefficient between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ has the formula

$$r^2_{y\hat{y}} = \frac{(\boldsymbol{y}'\hat{\boldsymbol{y}} - n\bar{y}^2)^2}{(\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2)(\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} - n\bar{y}^2)} \tag{1.2.10}$$

where $\bar{y}$ is the sample mean of $\boldsymbol{y}$ and also, note, the mean of $\hat{\boldsymbol{y}}$ by construction. This statistic, known as the *coefficient of determination* and more popularly denoted by R^2, is conventionally used to measure the goodness of fit of the regression, since it must lie between 0 and 1. It also has the more commonly quoted formulae

$$R^2 = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} - n\bar{y}^2}{\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2} \tag{1.2.11}$$

and, in view of (1.2.9),

$$R^2 = 1 - \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2}. \tag{1.2.12}$$

It can be verified that provided the regression includes an intercept, all three definitions are identical.

1.2.3 The Projection Matrices

The residual decomposition

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{1.2.13}$$

leads to the definition of two fundamental $n \times n$ matrices,

$$\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \tag{1.2.14}$$

and

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \tag{1.2.15}$$

such that $\hat{\boldsymbol{y}} = \boldsymbol{Q}\boldsymbol{y}$ and $\hat{\boldsymbol{u}} = \boldsymbol{M}\boldsymbol{y}$. Note that $\boldsymbol{Q}\boldsymbol{X} = \boldsymbol{X}$, and hence $\boldsymbol{M}\boldsymbol{X} = \boldsymbol{0}$. $\boldsymbol{Q}$ is called a *projection matrix.* In geometrical terms, think of it as projecting any n-vector into the space spanned by $\boldsymbol{X}$. $\boldsymbol{M}$ is the *orthogonal projection matrix*, projecting into the space orthogonal to $\boldsymbol{X}$. These matrices have the important properties of *symmetry:*

$$\boldsymbol{Q}' = \boldsymbol{Q}, \qquad \boldsymbol{M}' = \boldsymbol{M} \tag{1.2.16}$$

idempotency:

$$\boldsymbol{Q}\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{Q} \tag{1.2.17}$$

$$\boldsymbol{M}\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{Q})(\boldsymbol{I} - \boldsymbol{Q}) = \boldsymbol{I} - \boldsymbol{Q} - \boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{Q} = \boldsymbol{M} \tag{1.2.18}$$

and *mutual orthogonality:*

$$\boldsymbol{Q}\boldsymbol{M} = \boldsymbol{M}\boldsymbol{Q} = \boldsymbol{Q} - \boldsymbol{Q}\boldsymbol{Q} = \boldsymbol{0}. \tag{1.2.19}$$

Since the rank of an idempotent matrix is equal to its trace (see §A.6) it is easily shown that $\boldsymbol{Q}$ has rank k and $\boldsymbol{M}$ has rank $n - k$. Also note the relation

$$\hat{\boldsymbol{u}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{u} = \boldsymbol{M}\boldsymbol{u}. \tag{1.2.20}$$

One cannot of course use this to recover $\boldsymbol{u}$ from $\hat{\boldsymbol{u}}$ since $\boldsymbol{M}$ is singular, but it does yield the formula

$$S(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = \boldsymbol{u}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{M}\boldsymbol{u}. \tag{1.2.21}$$

1.2.4 Linear Transformations

Let $\boldsymbol{A}$ be a $k \times k$ nonsingular matrix, and define $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A}$. Each column of $\boldsymbol{Z}$ is a linear combination of columns of $\boldsymbol{X}$. The model can be written in terms of the new variables as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{u} \tag{1.2.22}$$

where $\boldsymbol{\delta} = \boldsymbol{A}^{-1}\boldsymbol{\beta}$. Note that the same transformation applies to the regression coefficients, since

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = (\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}} \tag{1.2.23}$$

where the final equality is obtained using (A.2.3). It follows directly that

$$\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\delta}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{u}} \tag{1.2.24}$$

so that an invertible linear transformation of the regressors leaves the residuals unchanged. The transformed model is statistically identical to the original, with only the interpretation of the coefficients changed.

1.2.5 The Partitioned Linear Model

Partition the regressors by columns as $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}$, where $\boldsymbol{X}_1$ is $n \times k_1$ and $\boldsymbol{X}_2$ is $n \times k_2$, where $k_1 + k_2 = k$, and let $\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix}\begin{matrix} k_1 \\ k_2\end{matrix}$ conformably. The regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ can then be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}. \tag{1.2.25}$$

It is frequently useful to have a formula for the k_1-dimensional sub-vector $\hat{\boldsymbol{\beta}}_1$, and this can be obtained by partitioning the normal equations $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ as

$$\begin{bmatrix}\boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y}\end{bmatrix} \tag{1.2.26}$$

or

$$\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_1'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1'\boldsymbol{y} \tag{1.2.27a}$$
$$\boldsymbol{X}_2'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\boldsymbol{y}. \tag{1.2.27b}$$

To solve these equations, first obtain $\hat{\boldsymbol{\beta}}_2$ from (1.2.27b), as

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'\boldsymbol{y} - \boldsymbol{X}_2'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1). \tag{1.2.28}$$

Substitution in (1.2.27a) and rearrangement yields

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= [\boldsymbol{X}_1'\boldsymbol{X}_1 - \boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1]^{-1}[\boldsymbol{X}_1'\boldsymbol{y} - \boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}] \\ &= (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y} \end{aligned}\tag{1.2.29}$$

where $\boldsymbol{M}_2 = \boldsymbol{I} - (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$. It may be verified that applying the partitioned inverse formula in (A.2.13) to the solution of (1.2.26) yields the same result.

Because of the idempotency and symmetry of $\boldsymbol{M}_2$ one can also write

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{M}_2'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2'\boldsymbol{y} \tag{1.2.30}$$

so that $\hat{\boldsymbol{\beta}}_1$ can be seen as the result of regressing $\boldsymbol{y}$ on $\boldsymbol{M}_2\boldsymbol{X}_1$. The latter is, in turn, the matrix of residuals from the regression of $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$. The result that multiple regression coefficients can be computed by this two-stage procedure is the *Frisch–Waugh theorem.* It was very useful in the days when regressions were computed by hand, see Frisch and Waugh (1933), and also Lovell (1963). Also note that replacing $\boldsymbol{y}$ by $\boldsymbol{M}_2\boldsymbol{y}$ in the formula leads to exactly the same result.

Another approach to the partitioned algebra is to apply an orthogonalizing transformation to the regressors. Define

$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{I}_{k_1} & \boldsymbol{0} \\ -(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{I}_{k_2}\end{bmatrix} \tag{1.2.31}$$

and so represent the model in the style of (1.2.22) as

$$\boldsymbol{y} = \boldsymbol{M}_2\boldsymbol{X}_1\boldsymbol{\delta}_1 + \boldsymbol{X}_2\boldsymbol{\delta}_2 + \boldsymbol{u} \tag{1.2.32}$$

where it is easily verified that $\boldsymbol{\delta}_1 = \boldsymbol{\beta}_1$ and $\boldsymbol{\delta}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2$. Since $\boldsymbol{X}_2'\boldsymbol{M}_2\boldsymbol{X}_1 = \mathbf{0}$, the partitioned inverse formula yields simply

$$\begin{bmatrix} \hat{\boldsymbol{\delta}}_1 \\ \hat{\boldsymbol{\delta}}_2 \end{bmatrix} = \begin{bmatrix} (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y} \\ (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix} \tag{1.2.33}$$

according to (1.2.23), where $\hat{\boldsymbol{\delta}}_1 = \hat{\boldsymbol{\beta}}_1$.

Finally, consider the residuals of the partitioned model. These are

$$\begin{aligned} \hat{\boldsymbol{u}} &= \boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 \\ &= \boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'\boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1) \\ &= \boldsymbol{M}_2\boldsymbol{y} - \boldsymbol{M}_2\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 \\ &= \left[\boldsymbol{M}_2 - \boldsymbol{M}_2\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\right]\boldsymbol{y} \end{aligned} \tag{1.2.34}$$

where the second equality substitutes from (1.2.28). However, since $\hat{\boldsymbol{u}} = \boldsymbol{M}\boldsymbol{y}$, and these equalities hold for arbitrary $\boldsymbol{y}$, it follows that

$$\boldsymbol{M} = \boldsymbol{M}_2 - \boldsymbol{M}_2\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2. \tag{1.2.35}$$

1.3 The Statistical Model

So much for computation. The next fundamental question concerns interpretation. Minimizing the sum of the squared prediction errors is not the only criterion for choosing an estimator of $\boldsymbol{\beta}$, so we would like to know how good an estimator $\hat{\boldsymbol{\beta}}$ is, and how it might compare with alternatives. To deal with these issues, some additional details about the observational setup, and the goals of the investigation, need to be filled in.

In the context of the linear model, an obvious question is the following: why in (1.1.1) is y_t placed on the left-hand side, with coefficient fixed at unity, and all the other variables on the right-hand side with unspecified coefficients? What is special about y_t? Two kinds of answer to this question have been offered above. The first relates to the purpose of the investigation. If this is prediction, in a situation where we observe (or have previously predicted) $\boldsymbol{x}_t$ and wish to predict y_t, an equation to predict y_t is what is needed. The second relates to the interpretation of the data generation mechanism. If y_t is the response of economic agents to the effects of $\boldsymbol{x}_t$, it is the natural choice for normalization because of the way the β_i are interpreted, as response coefficients.

This type of argument explains why the model is set up as it is, but it does not explain why least squares is the most suitable estimator. To do this requires a more formal approach. The set of assumptions about the way the sample data are generated go under the name of the *statistical model.* Broadly, three distinct statistical models of the data generation mechanism can be distinguished.

Model A: Fixed regressors

Elementary treatments of the regression model make the assumption that the regressors are *fixed in repeated samples.* This is basically a simplifying assumption to make the statistical analysis tractable, but it is as well to see just what it implies. This is the model most applicable to experimental situations. An experimenter chooses the experimental design, the cases $\boldsymbol{X} = \left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right]'$, and then $\boldsymbol{y} = (y_1, \ldots, y_n)'$ represents the experimental outcomes. In the regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, the $\boldsymbol{u}$ stands for errors in measuring $\boldsymbol{y}$, and other factors outside the experimenter's control. For example, consider an investigation of crop yields by agrobiologists. Let $\boldsymbol{y}$ represent the yields of a crop grown on n experimental plots, and let the rows of $\boldsymbol{X}$ represent the seed varieties, irrigation, fertilizer and other experimental variables for each plot. $\boldsymbol{u}$ represents the effects of uncontrolled differences from plot to plot, soil and sunlight variations for example, or errors in measuring the yield.

In this situation, the experiment can, in principle, be repeated as often as desired, with the same design matrix $\boldsymbol{X}$ but different, randomly drawn $\boldsymbol{u}$, and it is sensible to think of $\boldsymbol{X}$ as being fixed in repeated samples. The regression model is a natural basis for the analysis because only one of the variables, $\boldsymbol{y}$, is random.

Model B: Random regressors with independent sampling

Imagine a survey of household budgets, conducted by economists or government statisticians. A sample of families is drawn at random, and their incomes and expenditures on various commodities are recorded. An econometrician regresses (say) food expenditure on income and other relevant variables pertaining to consumption habits, things such as family size, number of children, etc. In this model the data are non-experimental, and the regressors cannot be described as fixed in repeated samples. If we draw a new sample of families, a new $\boldsymbol{y}$ *and* a new $\boldsymbol{X}$ are randomly selected each time.

However, the regression is still a valid device for predicting what a family with a given income etc. will spend on food, because it estimates the *conditional expectation* of y_t given $\boldsymbol{x}_t$. Thanks to the fact that the families are randomly sampled, it turns out that this model 'mimics' the fixed regressor model, and as we shall see in Chapter 2, many of the statistical properties of least squares in the fixed regressor model continue to hold. There are nonetheless some important differences from model A.

Model C: Time series regression (dependent sampling)

Suppose the data consist of annual or quarterly observations on national income and expenditures, drawn from the national accounts. In this model, the random sampling assumption does not apply. Successive periods in the history of the national economy are obviously highly dependent on one another, reflecting the business cycle, technological trends, and other factors carrying over from one year or quarter to the next. The same is true of nearly all economic data taking the form of a time series. When the sample points are dependent, the statistical model has *completely* different properties from cases such as randomly sampled households. It is fundamentally unlike the fixed regressor model. Special statistical treatment is called for here, although there are still important situations in which the least squares estimator is the correct tool of analysis.

This book will have little to say about model A, for the simple reason that economics (a basically non-experimental discipline) throws up little data of this type. In the present chapter, and Chapters 2 and 3, we are going to study model B in detail. Most of the remainder of the book will deal, implicitly or explicitly, with model C. In either model, the crucial concept in the analysis is going to be the conditional expectation. The formalities of conditioning theory are dealt with in Appendix B, but for present purposes, we will focus on intuition.

$E(y_t)$, the ordinary expected value of the random variable y_t, is usually to be thought of as the best predictor of y_t based on its behaviour in repeated sampling *and nothing else.* All the random influences that make one sample different from another are averaged out. On the other hand, in a sampling experiment generating the random vector $(y_t, \boldsymbol{x}_t)$, such as our sample survey, the conditional expectation $E(y_t|\boldsymbol{x}_t)$ can be thought of as the best predictor of y_t based on knowledge of $\boldsymbol{x}_t$. In essence, this is like a thought experiment. Compare the two questions: 'what does the average family spend on food?' and 'what does a family of two adults and two children living on £15,000 per annum expect to spend on food?'. The answers to both these questions can be expressed as mathematical expectations, but in the latter case it is a conditional expectation, corresponding to a thought experiment in which variables that actually vary randomly are given fixed chosen values.

Since $E(y_t|\boldsymbol{x}_t)$ is a function of $\boldsymbol{x}_t$, from the viewpoint of an observer to whom $\boldsymbol{x}_t$ is unknown it is itself a random variable. A fundamental property of conditional distributions is the *law of iterated expectations* (LIE), which states that the expectation of $E(y_t|\boldsymbol{x}_t)$ under the marginal distribution of $\boldsymbol{x}_t$ is identical with the expectation of y_t. That is,

$$E_x[E(y_t|\boldsymbol{x}_t)] = E(y_t) \tag{1.3.1}$$

where $E_x[\cdot]$ denotes the expectation under the marginal distribution.[2] (See Theorem B.6.1.)

In this context, constants can be thought of as a special variety of random variable, called *degenerate* (having zero variance). The intercept dummy and other 'non-random' elements of $\boldsymbol{x}_t$ are subsumed under this distribution. The linear regression model can now be understood as embodying the basic assumption,

$$\boldsymbol{x}_t'\boldsymbol{\beta} = E(y_t|\boldsymbol{x}_t) \tag{1.3.2}$$

or equivalently, $E(u_t|\boldsymbol{x}_t) = 0$.[3]

To show the implication of this assumption is very easy. Premultiply the expression by the vector $\boldsymbol{x}_t$ and take expectations through to give

$$E(\boldsymbol{x}_t\boldsymbol{x}_t')\boldsymbol{\beta} = E[\boldsymbol{x}_t E(y_t|\boldsymbol{x}_t)] = E[E(\boldsymbol{x}_t y_t|\boldsymbol{x}_t)] = E(\boldsymbol{x}_t y_t) \tag{1.3.3}$$

[2] Henceforth this subscript will usually be omitted, since the context should always make clear which variables are being 'averaged'.

[3] This statement of the model is technically imprecise, although the amendment required is generally of no practical importance. See §2.1 for an explanation, and also §B.6 and §B.10 for details.

where the second equality uses the fact that $\boldsymbol{x}_t$ can be treated 'like a constant' with respect to the distribution conditioned on it, and the last one is an application of the LIE. Assume that the matrix $E(\boldsymbol{x}_t\boldsymbol{x}_t')$ is positive definite, and hence nonsingular. This is equivalent to assuming that $E(\boldsymbol{x}_t'\boldsymbol{a})^2 > 0$ for an arbitrary vector of constants $\boldsymbol{a}$ (compare inequality (B.8.6)) and so does not do more than rule out any redundancy in the set of regressors. Then

$$\boldsymbol{\beta} = E(\boldsymbol{x}_t\boldsymbol{x}_t')^{-1}E(\boldsymbol{x}_t y_t) \tag{1.3.4}$$

is the expression that uniquely defines the coefficients of the linear conditional expectation of $y_t|\boldsymbol{x}_t$. There is an obvious resemblance to the formula for the OLS estimator. Indeed, replacing expectations by sample averages leads immediately to the formula

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{1.3.5}$$

noting that $\boldsymbol{X}'\boldsymbol{X} = \sum_{t=1}^{n} \boldsymbol{x}_t\boldsymbol{x}_t'$ and $\boldsymbol{X}'\boldsymbol{y} = \sum_{t=1}^{n} \boldsymbol{x}_t y_t$. Sums are equivalent to averages in this expression since factors of n^{-1} cancel.

1.4 Model Specification

1.4.1 Linearity

Formula (1.3.4) suggests the link between the least squares principle and a model of conditional expectations. However, the *linearity* in model (1.3.2) is no more than an assumption. For full generality, one should write

$$E(y_t|\boldsymbol{x}_t) = f_t(\boldsymbol{x}_t) \tag{1.4.1}$$

where $f_t(\cdot)$ is an arbitrary function, possibly depending on t. It may then be necessary to justify the linear function as an *approximation* to $f_t(\boldsymbol{x}_t)$. To see how linear regression should be interpreted in this case, consider the mean squared error function $E[E(y_t|\boldsymbol{x}_t) - \boldsymbol{x}_t'\boldsymbol{b}]^2$. One could think of the value of $\boldsymbol{b}$ that minimizes this expression as providing the 'best' linear approximation to $E(y_t|\boldsymbol{x}_t)$.

Theorem 1.4.1 If $E(\boldsymbol{x}_t\boldsymbol{x}_t')$ is positive definite, the mean squared approximation error is minimized by setting $\boldsymbol{b} = \boldsymbol{\beta}_t$ where

$$\boldsymbol{\beta}_t = [E(\boldsymbol{x}_t\boldsymbol{x}_t')]^{-1}E(\boldsymbol{x}_t y_t). \tag{1.4.2}$$

Proof Substituting and rearranging yields

$$\begin{aligned} E(E(y_t|\boldsymbol{x}_t) - \boldsymbol{x}_t'\boldsymbol{b})^2 &= E[E(y_t|\boldsymbol{x}_t) - \boldsymbol{x}_t'\boldsymbol{\beta}_t + \boldsymbol{x}_t'(\boldsymbol{\beta}_t - \boldsymbol{b})]^2 \\ &= E[E(y_t|\boldsymbol{x}_t) - \boldsymbol{x}_t'\boldsymbol{\beta}_t]^2 + (\boldsymbol{\beta}_t - \boldsymbol{b})'E(\boldsymbol{x}_t\boldsymbol{x}_t')(\boldsymbol{\beta}_t - \boldsymbol{b}) \\ &\quad + 2(\boldsymbol{\beta}_t - \boldsymbol{b})'E[\boldsymbol{x}_t(E(y_t|\boldsymbol{x}_t) - \boldsymbol{x}_t'\boldsymbol{\beta}_t)]. \end{aligned} \tag{1.4.3}$$

However, since $E[\boldsymbol{x}_t E(y_t|\boldsymbol{x}_t)] = E(\boldsymbol{x}_t y_t)$ by the LIE,

$$E[\boldsymbol{x}_t(E(y_t|\boldsymbol{x}_t) - \boldsymbol{\beta}_t'\boldsymbol{x}_t)] = E(\boldsymbol{x}_t y_t) - \boldsymbol{\beta}_t'E(\boldsymbol{x}_t\boldsymbol{x}_t') = \boldsymbol{0}. \tag{1.4.4}$$

Hence, (1.4.3) is minimized by setting $\boldsymbol{b} = \boldsymbol{\beta}_t$, by Lemma A.7.3. ■

It further follows that for any given $\boldsymbol{x}_t$, $\boldsymbol{\beta}_t'\boldsymbol{x}_t$ is the linear predictor that minimizes the mean squared error in predicting y_t from $\boldsymbol{x}_t$.

Theorem 1.4.2 The function $E(y_t - \boldsymbol{b}'\boldsymbol{x}_t)^2$ is minimized by setting $\boldsymbol{b} = \boldsymbol{\beta}_t$.

Proof

$$\begin{aligned} E(y_t - \boldsymbol{b}'\boldsymbol{x}_t)^2 &= E[(y_t - E(y_t|\boldsymbol{x}_t) + E(y_t|\boldsymbol{x}_t) - \boldsymbol{b}'\boldsymbol{x}_t]^2 \\ &= E[y_t - E(y_t|\boldsymbol{x}_t)]^2 + E[E(y_t|\boldsymbol{x}_t) - \boldsymbol{b}'\boldsymbol{x}_t]^2 \\ &\quad + 2E[(y_t - E(y_t|\boldsymbol{x}_t))(E(y_t|\boldsymbol{x}_t) - \boldsymbol{b}'\boldsymbol{x}_t)]. \end{aligned} \tag{1.4.5}$$

The cross-product term vanishes, since by the LIE

$$\begin{aligned} E[(y_t - E(y_t|\boldsymbol{x}_t))(E(y_t|\boldsymbol{x}_t) - \boldsymbol{b}'\boldsymbol{x}_t)] &= E[(E(y_t|\boldsymbol{x}_t) - E(y_t|\boldsymbol{x}_t))(E(y_t|\boldsymbol{x}_t) - \boldsymbol{b}'\boldsymbol{x}_t)] \\ &= E(0) = 0. \end{aligned} \tag{1.4.6}$$

The result therefore follows by Theorem 1.4.1. ■

Equation (1.4.2) differs from (1.3.4) since it is not necessarily the case that $\boldsymbol{\beta}_t = \boldsymbol{\beta}$, independent of t. This would be true if, for example, the data were identically distributed, implying in particular that $E(\boldsymbol{x}_t\boldsymbol{x}_t')$ and $E(\boldsymbol{x}_t y_t)$ are constants which do not depend on t. Later on, we show that $\hat{\boldsymbol{\beta}}$, more generally, estimates a weighted average of the $\boldsymbol{\beta}_t$,

$$\boldsymbol{\beta}_n = \left(\sum_{t=1}^n E(\boldsymbol{x}_t\boldsymbol{x}_t')\right)^{-1} \sum_{t=1}^n E(\boldsymbol{x}_t y_t) = \sum_{t=1}^n \boldsymbol{A}_{nt}\boldsymbol{\beta}_t \tag{1.4.7}$$

where

$$\boldsymbol{A}_{nt} = \left(\sum_{t=1}^n E(\boldsymbol{x}_s\boldsymbol{x}_s')\right)^{-1} E(\boldsymbol{x}_t\boldsymbol{x}_t') \tag{1.4.8}$$

having the property $\sum_{t=1}^n \boldsymbol{A}_{nt} = \boldsymbol{I}_k$. The term 'estimate' is used here in the rather special sense that $\boldsymbol{\beta}_n$ and the least squares estimator are converging to the same limit (appropriately defined, and assuming such limits exist) as n tends to infinity. The technical details are covered in §10.5.

Unless stated specifically to the contrary, in what follows we will always appeal to the so-called *axiom of correct specification*, or in other words assume (1.3.2) is true. The properties of the OLS estimator to be derived in Chapter 2 and subsequently do not generally hold without (1.3.2), but these remarks may serve to reassure us that the assumption need not be too critical.

1.4.2 Included Variables

Even if linearity is assumed, our model may be *incomplete* in the following sense. Suppose there are some other variables $\boldsymbol{z}_t$ $(l \times 1)$ relevant to the explanation of y_t, such that

$$E(y_t|\boldsymbol{x}_t, \boldsymbol{z}_t) = \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{z}_t'\boldsymbol{\delta}. \tag{1.4.9}$$

Then, evidently,

$$E(y_t|\boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\beta} + E(\boldsymbol{z}_t'|\boldsymbol{x}_t)\boldsymbol{\delta}. \tag{1.4.10}$$

In general, the second term is an arbitrary function of $\boldsymbol{x}_t$. However, if the linearity assumption is extended by assuming that $E(\boldsymbol{z}_t|\boldsymbol{x}_t) = \boldsymbol{D}\boldsymbol{x}_t$ where

$$\boldsymbol{D} = E(\boldsymbol{z}_t\boldsymbol{x}_t')E(\boldsymbol{x}_t\boldsymbol{x}_t')^{-1} \quad (l \times k) \tag{1.4.11}$$

is independent of t, then

$$E(y_t|\boldsymbol{x}_t) = \boldsymbol{x}_t'\boldsymbol{\gamma} \tag{1.4.12}$$

where $\boldsymbol{\gamma} = \boldsymbol{\beta} + \boldsymbol{D}'\boldsymbol{\delta}$. The vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are called respectively the *simple* and *partial* regression coefficients of $\boldsymbol{x}_t$. The least squares coefficients have a corresponding decomposition. Note that the formula (1.2.28) applied to the regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{u} \tag{1.4.13}$$

yields $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{D}}'\hat{\boldsymbol{\delta}}$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$ are the multiple regression coefficients, $\hat{\boldsymbol{\gamma}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and $\hat{\boldsymbol{D}} = \boldsymbol{Z}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. These formulae reveal how the computational and the interpretive relationships between the simple and partial coefficients are closely linked.

It is possible that $\boldsymbol{\gamma}$ represents the parameters of interest in the investigation. For example, $\boldsymbol{x}_t'\boldsymbol{\gamma}$ is the best (minimum mean squared error) predictor of y_t given $\boldsymbol{x}_t$, and if $\boldsymbol{z}_t$ were unobserved, prediction by $\boldsymbol{x}_t$ might be the object of the exercise so $\boldsymbol{\gamma}$ is what we wish to know. However, $\boldsymbol{\beta}$ is more commonly regarded as containing the parameters of interest, because they show how y_t and $\boldsymbol{x}_t$ are directly related, after controlling for the effects of $\boldsymbol{z}_t$. They are commonly identified with theoretical economic magnitudes (e.g. elasticities of demand or supply, technological coefficients) whereas no such interpretation is available for composite parameters such as $\boldsymbol{\gamma}$. Moreover, the assumption $E(\boldsymbol{z}_t|\boldsymbol{x}_t) = \boldsymbol{D}\boldsymbol{x}_t$ is entirely *ad hoc*. If in reality $E(\boldsymbol{z}_t\boldsymbol{x}_t')E(\boldsymbol{x}_t\boldsymbol{x}_t')^{-1} = \boldsymbol{D}_t$, depending on t, the composite coefficients are $\boldsymbol{\gamma}_t = \boldsymbol{\beta} + \boldsymbol{D}_t'\boldsymbol{\delta}$ and linearity fails for the incomplete model even if it holds for the complete one. Therefore, unless $\boldsymbol{z}_t$ is included in the regression, the model is misspecified.

However, in econometrics, the number of relevant variables may be very large, often exceeding the number of observations available, and choices of a practical nature are forced on the investigator. If the elements of $\boldsymbol{\delta}$ are of no interest, so-called *nuisance* parameters, and also $\boldsymbol{\beta} = \boldsymbol{\gamma}$, a case can often be made for excluding $\boldsymbol{z}_t$. Note that the latter condition holds if $\boldsymbol{D} = \boldsymbol{0}$, or more generally if $E(\boldsymbol{z}_t|\boldsymbol{x}_t) = \boldsymbol{0}$,[4] and need not imply $\boldsymbol{\delta} = \boldsymbol{0}$. In principle, it would be preferable to include the variables in every case where $\boldsymbol{\delta} \neq \boldsymbol{0}$, because otherwise $\boldsymbol{z}_t'\boldsymbol{\delta}$ becomes

[4] This is a mild abuse of notation since strictly, when $\boldsymbol{z}_t$ is not predictable from $\boldsymbol{x}_t$ the conditional mean is a constant, not necessarily zero. Similarly, if $x_{kt} = 1$, the kth column of $\boldsymbol{D}$ should not vanish. However, this term can be subsumed by adding it to the intercept of the regression, and it can be conveniently neglected by assuming the variables are expressed in deviations from the mean.

a component of the error term. Even if $\boldsymbol{\beta} = \boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}$ is an inferior estimator of $\boldsymbol{\beta}$ relative to $\hat{\boldsymbol{\beta}}$, other things equal, because the error term has a larger variance in that case. However, if $\boldsymbol{\delta} = \mathbf{0}$ while $\boldsymbol{D} \neq \mathbf{0}$, it is also known that $\hat{\boldsymbol{\beta}}$ is typically inferior to $\hat{\boldsymbol{\gamma}}$.[5] The case where $\boldsymbol{\delta}$ is nonzero but 'small', in some suitable sense, may therefore represent a balance of advantage between inclusion and exclusion. Given that prior knowledge of all these relations is inevitably lacking in practice, there can be no simple rule underlying the choice of regressors. A good deal of our subsequent analysis will focus on this problem.

1.4.3 Relevant Variables

We neglected an important and difficult question in the previous section by speaking of the *relevant* variables in the model. A distinction needs to be made is between *explanatory variables*, and what may be called the *valid conditioning variables.* To be relevant, a variable must fall into both of these categories. Suppose the object of the exercise is to forecast y_t in a situation where $\boldsymbol{x}_t$ is observed, but y_t and also $\boldsymbol{z}_t$ are unobserved. Then, no matter how 'good' an equation (1.4.9) may be in terms of least squares fit, it is of no interest. Trivially, putting $\boldsymbol{z}_t = y_t$ one could write $E(y_t|\boldsymbol{x}_t, y_t) = y_t$ and obtain a model that forecasts perfectly, but is useless. For this problem, y_t is not a valid conditioning variable.

Similarly, a behavioural model must make economic sense. It is possible to 'explain' a families' total income in terms of its food expenditure, in the sense that these tend to move together in a predictable fashion, and either can be used to predict the other. However, we aren't usually interested in the regression of income on expenditure. This is because a model of economic behaviour based on the conditional expectation $E(y_t|\boldsymbol{x}_t)$ embodies an assumption about the order of causation. $\boldsymbol{x}_t$ must be given (i.e., observed by agents) at the moment when y_t is determined, such that no feedback from y_t to $\boldsymbol{x}_t$ is possible. If, in a model of economic behaviour, the disturbance term represents the deviation between plans and outcomes due to unforeseen events, it is appropriate for it to be unpredictable by the conditioning variables, in the sense $E(u_t|\boldsymbol{x}_t) = 0$.

In the context of the food demand example, a valid conditioning variable is any variable that the household has observed when the spending decision is made. It is also an explanatory variable if it actually influences their decision. The expenditure model would not be valid if households were to plan *jointly* what to earn and what to spend on food, so that the two decisions are interdependent rather than sequential. Over a longer time-scale (say, lifetimes rather than months or years), we would typically recognize the existence of such an interdependence. In general, the simple regression model is appropriate to cases where a single variable is chosen conditional on others. The alternative possibility, with two or more variables being jointly determined, is a model involving several simultaneous equations. This type of model is studied in Chapters 8 and 13.

[5] The statistical arguments underlying these claims are explored in Chapter 2 and subsequently.

Further Reading: Alternative accounts of the linear regression algebra can be found in many popular textbooks. Theil (1971) is recommended. Rao (1973), Madansky (1976), Seber (1980) offer good advanced treatments. For a geometric interpretation of the algebra, see Davidson and MacKinnon (1993). Goldberger (1991) and Spanos (1986) emphasize the 'conditional expectation' interpretation. On nonlinear regression, see Gallant (1987), Davidson and MacKinnon (1993), Malinvaud (1970).