Part II

Functional Forms in Production Theory

Chapter II.1

A SURVEY OF FUNCTIONAL FORMS IN THE ECONOMIC ANALYSIS OF PRODUCTION

MELVYN FUSS, DANIEL McFADDEN, and YAIR MUNDLAK*

University of Toronto, University of California, Berkeley, and
The Hebrew University of Jerusalem

1. The Context and Objectives of Production Analysis

1.1. *Introduction*

Empirical analysis of technology is carried out in many contexts, for many purposes. Each situation raises specific conditions and objectives which must be met in the specification of an econometric production model. This chapter surveys a variety of functional forms for production processes, and their cost and profit duals, and discusses some of the applications for which they are suited.

The diversity and extent of the subject of applied production theory makes a comprehensive survey impossible. We emphasize the structure of alternative functional forms, and the relationship between "exact" models of technology and econometric models incorporating stochastic specifications. However, we have not attempted to provide either a full catalog of properties of functional forms or a general procedure for introducing stochastic elements in production models. We focus on several basic issues of technology - scale, separability, and substitutability. We have not attempted to treat a number of other major issues,

*The financial support through NSF grant no. SOC-73-05374-A01 is gratefully acknowledged by Y. Mundlak.

such as technical change and aggregation, which are equally important in many applications.

The uses of production models can be classified in two ways. The first is the distinction between analytic studies of the production process (for example, a test of the constancy of returns to scale), versus estimation to provide predictions for specific applications (for example, a prediction of industrial demand for energy). The former alternative requires close attention to the structure and parameterization of the production model, while the latter is more concerned with the robustness of the model and its extrapolative plausibility.

The second division is between macroeconomic analysis of aggregative production relationships and microeconomic treatment of industry, firm, and establishment technologies. Issues of aggregation over commodities, economic units, and technologies, and questions of proper parameterization of distribution, technical change, and growth effects have dominated the literature on aggregate production functions. Questions of compatibility with physical production processes and firm behavior have been important in the analysis of microeconomic production relations. Statistical issues in the estimation of technological relationships have concentrated on the stochastic nature of aggregate quantity and price indices, as determined by their definition and measurement, and on the stochastic specification of microeconomic firm behavior.

In a survey of functional forms, it is important to keep in mind the fact that these forms have been constructed for a variety of applications. One cannot expect to find a single "best" parametric production function for all purposes; to the contrary, many of these functional forms are well-suited for specific applications but poorly-suited for use as general purpose characterizations of technology.

1.2. *Objectives of Production Analysis*

Historically, emphasis has been placed on a number of different aspects of technology, depending on the objectives of analysis. We list below some of the major objectives of production studies which have motivated the development of functional forms:

(1) *Distribution* (the income shares of factors of production): Most attention has centered on the aggregate shares of capital and labor. Distribution issues also arise at the microeconomic level in problems

such as the incidence of tax and subsidy programs. Distribution parameters are of great importance in evaluating the growth process.

(2) *Scale* (the existence of constant returns to scale, or the presence of decreasing or increasing returns): Scale has aggregate implications for long-run growth, and for the structure of industry, which is also related to the question of the logical consistency of the neoclassical assumption of profit maximization. Microeconomic issues which focus on the supply and financing of public services often center on the technological question of the existence of increasing returns to scale.

(3) *Substitution* (the degree of substitutability of factors of production): Substitutability is a critical issue in the behavior of distributive shares when factors proportions change. It plays an important role in determining the incidence of taxes; and also the behavior of relative factor prices, and therefore product prices, in the process of growth.

(4) *Separability* (decomposition of production relationships into nested or additive components): Separability is an extremely important structural property in a production model which often permits econometric analysis to be carried out in terms of subsets of the total set of possible variables, in stages, or with consistent aggregates of variables. Separability is of direct economic interest, implying uniform or invariant behavior of certain economic quantities, and allowing decentralization in decision-making. It is also of critical interest in the specification of functional forms, influencing generality and simplicity, and becomes an important subject for empirical tests. (Because of its pivotal role in functional form specification, separability is discussed in detail in Section 6 of this survey.)

(5) *Technical change* (modification of the technological structure over time): Of interest are disembodied technical change (innovations which require no specific capital); technical change embodied in factors of production (usually capital, but potentially other factors such as skilled labor); factor-augmenting change which increases the effective quality of inputs; augmentation of other technological characteristics such as scale-augmenting change (increasing the scale level at which decreasing returns set in) or substitution-augmenting change (increasing the substitutability of inputs); and endogenous technical change (learning-by-doing; innovation and induced technical change).

In addition, a number of auxiliary topics have been the subject of econometric investigation, with attendant problems of functional specification.

1. *Technological flexibility* (the robustness of the technology in adapt-

ing to changing environments): Of interest is the degree to which flexibility is incorporated in the adopted technology, and its tradeoff against static efficiency.

2. *Efficiency* (operation on or inside the technology frontier): Relative efficiency of different economic units (firms, industries, nations) is of interest, as is the efficiency of the same unit in alternative economic environments.

3. *Homotheticity* (the presence of expansion paths with scale which are rays through the origin): Homothetic production functions will display unchanging distributive shares with changes in scale, ceteris paribus. In contrast, heterotheticity will yield changing factor intensities with changes in scale.

4. *Consistent aggregation* (the problem of specifying technological structures that are invariant with respect to aggregation over commodities or economic units): This problem is most critical in studies which want to ensure microeconomic compatibility of aggregate analysis, or want to obtain simple aggregate forecasts from microeconomic estimates.

In surveying various forms, one should keep in mind the alternative objectives listed above.

2. Criteria for the Design of Functional Forms

2.1. *Maintained Hypotheses*

In addition to the obvious criterion that a functional form should relate to the objectives of an analysis, there are a few general principles which should be adopted in modelling production. The first concerns the role of maintained hypotheses.

Any study in production economics (and, for that matter, in econometrics in general) takes place against the background of a series of maintained hypotheses which are not themselves tested as part of the analysis, but are assumed true. The most fundamental of these maintained hypotheses are basic axioms on the nature of technology (e.g., "the production possibility set is closed"), which are widely accepted because they are believed to be true, or at least irrefutable with existing data. Second come technological and behavioral assumptions which are not widely held to be universal truths, but may be widely accepted as plausible for the problem at hand (e.g., "convex technology" or "cost minimizing behavior"). Next come assumptions made to facilitate the analysis (e.g., "independent normal errors", "intermediate inputs separable from primary inputs"), which are believed to be harmless approximations to reality. Finally, there may be maintained hypotheses, such as the assumption of a specific parametric functional form, or of the constancy of some unobserved prices or quantities, which are accepted only for convenience or tractability. The analyst may then argue that his results are robust or insensitive with respect to these hypotheses, justifying their imposition on grounds of usefulness and lack of negative consequence rather than on grounds of plausibility.

The outcome of a specific test of hypothesis depends in general on *both* the validity of the hypothesis under examination and the validity of the maintained hypotheses. Consequently, a test performed in the presence of an implausible maintained hypothesis may not be convincing; the result may be a consequence of the validity of the maintained hypothesis rather than of the primary hypothesis in which one is interested. This suggests the general principle that *one should not attempt to test a hypothesis in the presence of maintained hypotheses that have less commonly accepted validity.* For example, it would be inappropriate to test a basic assumption such as convexity of the technology by examining the sign of the estimated elasticity of substitution when a C.E.S. production function is imposed as a maintained hypothesis, since a rejection is more likely to be interpreted as a failure of the C.E.S. specification than of convexity. An implication of this principle is the need for general, flexible functional forms, embodying few maintained hypotheses, to be used in tests of the fundamental hypotheses of production theory. Given the qualitative, non-parametric nature of the fundamental axioms, this suggests further that the more relevant tests will be non-parametric, rather than based on parametric functional forms, even very general ones. While non-parametric approaches to the study of production relationships have received some attention in economics [Farrell (1957), Hanoch and Rothschild (1972)], these methods have been exploited less systematically for tests of basic hypotheses than have parametric forms [e.g. Berndt-Christensen (1973a)]. Analyses of the latter type inevitably are subject to the criticism that a rejection of a hypothesis may be a result of the parametric specification rather than falseness of the hypothesis. This criticism must be balanced, however, against the observation that non-parametric tests have not yet been developed for some multivariate production hypotheses.

For most analyses, the econometrician has a choice of several starting points for the specification of functional forms. This book emphasizes the equivalence of production, cost, and profit functions as characterizations of technology under appropriate conditions (including competitive markets). It is also possible to specify a production model directly in terms of demand and supply functions, expressed either in prices or quantities, or even in terms of differential or difference equations for these demand and supply functions. Under appropriate integrability conditions, these systems can then be solved to obtain the implied production, cost, or profit functions. This survey will emphasize functional forms for production, cost, and profit functions, but will not attempt to survey specifications of technology which are formulated directly in terms of demand and supply functions or their derivatives.

2.2. *Criteria for Choosing Functional Forms*

Within the framework of the maintained hypotheses imposed on a particular problem or class of problems, a wide variety of compatible functional forms will usually be available. We list some of the criteria which may be used to select among them:

1. *Parsimony in parameters*: The functional form should contain no more parameters than are necessary for consistency with the maintained hypotheses. Excess parameters exacerbate problems of multicollinearity, which tend to be severe in any case in many applications due to market substitution which causes prices, and hence quantities, to be highly correlated. Further, when the sample is small, excess parameters mean a loss of degrees of freedom, a particular problem in aggregate analysis.

2. *Ease of interpretation*: Excessively complex or parameter-rich functional forms may contain implausible implications which are hidden from easy detection. Further, complex transformations may make it cumbersome to compute and assess economic effects of interest; for example, elasticities of substitution. Thus, ceteris paribus, it is better to choose functional forms in which the parameters have an intrinsic and intuitive economic interpretation, and in which functional structure is clear.

3. *Computational ease*: Historically, systematic multivariate empirical analysis has been confined to linear (in parameters) statistical models for computational reasons. While current computational technology makes

direct estimation of non-linear forms feasible, it remains the case that linear-in-parameters systems have a computation cost advantage, and have, in addition, the advantage of a more fully-developed statistical theory. The tradeoff between the computational requirements of a functional form and the thoroughness of empirical analysis should be weighed carefully in the choice of a model.

4. *Interpolative robustness*: Within the range of observed data, the chosen functional form should be well-behaved, displaying consistency with maintained hypotheses such as positive marginal products or convexity. If these properties must be checked numerically, then the form should admit convenient computational procedures for this purpose.

5. *Extrapolative robustness*: The functional form should be compatible with maintained hypotheses *outside* the range of observed data. This is a particularly important criterion for forecasting applications.

3. Dual Transformation, Cost, and Profit Functions – Maintained Hypotheses on the Technology and Its Representations[1]

In this section, we summarize the commonly imposed maintained hypotheses for production, cost, and profit functions. Much of the development of specific functional forms has concentrated on questions of consistency with these hypotheses. More detailed discussions of the relationships among these properties are given in Part I of this volume.

3.1. *Production Possibility and Input Requirement Sets*

The basic notion to be introduced is that of a technology. Let $\mathbf{v},\mathbf{y}$ be vectors of inputs and outputs, respectively. The *production possibility set* $\mathbf{Y}$ is the set of all feasible input–output combinations, i.e., $\mathbf{Y} = \{\mathbf{v},\mathbf{y}:\mathbf{v}$ can yield $\mathbf{y}\}$. For each $\mathbf{y}$ occurring in some input–output vector in $\mathbf{Y}$ we can define the *input requirement set* $\mathbf{V}(\mathbf{y})$, containing all the input bundles which can produce $\mathbf{y}$, i.e., $\mathbf{V}(\mathbf{y}) = \{\mathbf{v}:(\mathbf{v},\mathbf{y}) \in \mathbf{Y}\}$. It is convenient to describe the maintained hypotheses on the technology in terms of the properties of $\mathbf{V}(\mathbf{y})$.

[1] This section is intended as a summary in order to make the chapter self-contained. A more detailed description of the characteristics of the representations of technology can be found in Chapter I.1 of this volume.

The properties of V(y) are assumed to be:

1.1 *Location.* V(y) is a non-empty subset of the non-negative orthant $\mathbf{R}^n$, denoted by $\mathbf{\Omega}_n$. It is possible that some factors will not be utilized. However, the only output that can be obtained with no inputs at all is the zero output. It is therefore required that $\mathbf{V(0)} = \mathbf{\Omega}_n$ and $\mathbf{y} > \mathbf{0}$ imply $\mathbf{0} \notin \mathbf{V(y)}$.

1.2 *Closure.* The analysis is greatly simplified when V(y) is assumed to be closed. That is, if a sequence of points $\{\mathbf{v}^n\}$ in V(y) converges, the limiting point also belongs to V(y). This means that V(y) contains all its limit points, and assures that the efficiency frontier of V(y) belongs to V(y).

1.3 *Monotonicity.* If a given output can be produced by the input-mix **v** it can also be produced by a larger input: if $\mathbf{v} \in \mathbf{V(y)}$ and $\mathbf{v}' \geq \mathbf{v}$ then $\mathbf{v}' \in \mathbf{V(y)}$. Similarly, the inputs required to produce a given output can certainly produce a smaller output. If $\mathbf{y} \geq \mathbf{y}'$ then $\mathbf{V(y)} \subset \mathbf{V(y')}$. These conditions imply that, unless **Y** is bounded and the boundary belongs to **Y**, there is no input-mix that can produce every **y** in **Y**.

1.4 *Convexity.* V(y) is convex.

3.2. *Production and Distance Functions*

Suppose we restrict **y** to a single element *y*. Then, using the notion of the input requirement set, the *production function* for *y* can be defined by

$$f(\mathbf{v}) = \max_{y} \{y : \mathbf{v} \in \mathbf{V}(y)\}.$$

When **V**(*y*) has properties (1.1) through (1.4), *f*(**v**) has the following properties (Diewert (1971)):

2.1 *Domain.* *f*(**v**) is a real-valued function of **v** defined for every $\mathbf{v} \in \mathbf{\Omega}_n$ and it is finite if **v** is finite; $f(\mathbf{0}) = 0$.

2.2 *Monotonicity.* An increase in inputs cannot decrease production:

$$\mathbf{v} \geq \mathbf{v}' \Rightarrow f(\mathbf{v}) \geq f(\mathbf{v}').$$

2.3 *Continuity.* f is continuous from above: every sequence $\{\mathbf{v}^n\} \subset \mathbf{\Omega}_n$ such that $f(\mathbf{v}^n) \geq y^0$, $y^0 = f(\mathbf{v}^0)$ and $\mathbf{v}^n \to \mathbf{v}^0$ implies $\lim_{n\to\infty} f(\mathbf{v}^n) = y^0$. Of course, this is a weaker property than continuity, which is almost universally imposed on the production function in empirical work.

2.4 *Concavity.* f is quasi-concave over $\mathbf{\Omega}_n$: the set $\{\mathbf{v}: f(\mathbf{v}) \geq y, \mathbf{v} \in \mathbf{\Omega}_n\}$ is convex for every $y \geq 0$. This property insures diminishing marginal rates of substitution.

In addition, twice differentiability of f is commonly imposed in empirical work.

When $\mathbf{y}$ contains more than one element, efficient production of $\mathbf{y}$ can be described in terms of the *distance*[2] *function*

$$D(\mathbf{y},\mathbf{v}) = \max\left\{\lambda > 0 \,\middle|\, \frac{1}{\lambda}\mathbf{v} \in \mathbf{V}(\mathbf{y})\right\},$$

for $(\mathbf{v},\mathbf{y}) \in \mathbf{Y}$ and $\mathbf{v}$ strictly positive; the frontier satisfies $D(\mathbf{y},\mathbf{v}) = 1$.

Alternatively we can define the transformation function as the maximum amount of y_1 which can be produced given the amounts of the other commodities $\hat{\mathbf{y}} = (y_2,\ldots,y_n)$ and $\mathbf{v} = (v_1,\ldots,v_n)$, i.e.,

$$F(\hat{\mathbf{y}},\mathbf{v}) = \max_{y_1}\{y_1 : (y_1,\hat{\mathbf{y}},\mathbf{v}) \in \mathbf{Y}\}.$$

The transformation function is assumed to have the following properties [Diewert (1974a)]:

2.1.1 *Domain.* F is an extended real-valued function defined and bounded from above for every $(\hat{\mathbf{y}},\mathbf{v}) \in \mathbf{\Omega}_{n+m-1}$. Also,

$$F(\mathbf{0},\mathbf{0}) = 0.$$

2.1.2 *Monotonicity.* F is non-increasing in $\hat{\mathbf{y}}$ and non-decreasing in $\mathbf{v}$.

2.1.3 *Continuity.* F is continuous from above.

2.1.4 *Concavity.* F is a concave function.

[2] For a detailed description of distance functions, see Chapter I.1.

The distance function and transformation function have a simple relationship:

$$D(\mathbf{y},\mathbf{v}) = \max\{\lambda > 0 | y_1 = F(\hat{\mathbf{y}},\mathbf{v}/\lambda)\},$$

and $y_1 = F(\hat{\mathbf{y}},\mathbf{v})$ is the solution to the equation

$$D(y_1,\hat{\mathbf{y}},\mathbf{v}) = 1.$$

Then, $D(F(\hat{\mathbf{y}},\mathbf{v}),\hat{\mathbf{y}},\mathbf{v}) \equiv 1$ is an identity, as is $y_1 \equiv F(\hat{\mathbf{y}},\mathbf{v}/D(\mathbf{y},\mathbf{v}))$. Using the properties of distance functions derived in Chapters I.1 and I.3, the reader can use these identities to deduce the properties of the transformation function.

3.3. *The Cost Function*

In general, economic models involving production need, in addition to the production function or transformation function, rules of behavior. The selection of the optimal input mix for some $\mathbf{y} \in \mathbf{Y}$ and some set of exogenous input prices $\mathbf{r}$ normally assumes cost minimizing behavior. Cost minimization for all $\mathbf{r} \in \mathbf{\Omega}_n^*$, where $\mathbf{\Omega}_n^*$ is the strictly positive orthant, and $\mathbf{y} \in \mathbf{Y}$ is described by the cost function

$$C(\mathbf{y},\mathbf{r}) = \min\{\mathbf{r}\cdot\mathbf{v}:\mathbf{v} \in \mathbf{V}(\mathbf{y})\}.$$

If the input markets are not competitive, a cost function can still be defined by this formula, with the prices $\mathbf{r}$ interpreted as shadow or imputed prices.

If $\mathbf{V}(\mathbf{y})$ possesses Properties (1.1) through (1.4) then $C(\mathbf{y},\mathbf{r})$,has Properties (3.1)–(3.5) listed below:

3.1 *Domain.* $C(\mathbf{y},\mathbf{r})$ is a positive real-valued function defined for all positive prices $\mathbf{r}$ and all positive producible outputs; $C(\mathbf{0},\mathbf{r}) = 0$.

3.2 *Monotonicity.* $C(\mathbf{y},\mathbf{r})$ is a non-decreasing function in output and tends to infinity as output tends to infinity. It is also non-decreasing in prices.

3.3 *Continuity.* $C(\mathbf{y},\mathbf{r})$ is continuous from below in $\mathbf{y}$ and continuous in $\mathbf{r}$.

3.4 *Concavity.* $C(\mathbf{y},\mathbf{r})$ is a concave function in $\mathbf{r}$.

3.5 *Homogeneity.* $C(\mathbf{y},\mathbf{r})$ is linear homogeneous in $\mathbf{r}$.

Empirical work usually assumes in addition:

3.6 *Differentiability.* In most empirical applications, $C(\mathbf{y},\mathbf{r})$ is to be twice differentiable in $\mathbf{r}$.

Under 3.6, the cost function possesses the important derivative property

(a) $\frac{\partial C}{\partial r_i} = v_i$ (Shephard's Lemma);

from which it follows that

(b) $\frac{\partial^2 C}{\partial r_i \partial r_j} = \frac{\partial^2 C}{\partial r_j \partial r_i}$ or $\frac{\partial v_i}{\partial r_j} = \frac{\partial v_j}{\partial r_i}$ (Symmetry).

Property (a) can be used to generate systems of factor demand functions. Property (b) is of use in reducing the number of parameters to be estimated, thus conserving degrees of freedom and possibly eliminating multicollinearity problems.

3.4. *The Profit Function*

Cost minimization can be construed as the first stage of a two-stage procedure. The second stage, given an exogenous output price vector $\mathbf{s}$, is the selection of $\mathbf{y}$ to maximize profit. Profit maximization for all $\mathbf{r} \in \mathbf{\Omega}_n^*$, $\mathbf{s} \in \mathbf{\Omega}_m^*$ is described by the profit function

$$\Pi(\mathbf{s},\mathbf{r}) = \max\{\mathbf{s}\cdot\mathbf{y} - \mathbf{r}\cdot\mathbf{v}:(\mathbf{v},\mathbf{y}) \in \mathbf{Y})\},$$

or

$$\Pi(\mathbf{p}) = \max\{\mathbf{p}\cdot\mathbf{x}:\mathbf{x} \in \mathbf{Y}\},$$

where $\mathbf{x}$ is a net output vector (>0 for outputs, and <0 for inputs) and $\mathbf{p} = (\mathbf{r},\mathbf{s}) \in \mathbf{\Omega}_{m+n}^*$.

If $F(\hat{\mathbf{y}},\mathbf{v})$ possesses Properties (2.1.1) through (2.1.4) then $\Pi(\mathbf{p})$ has Properties (4.1)–(4.5):

4.1 *Domain.* $\Pi(\mathbf{p})$ is a non-negative extended real-valued function defined for all positive prices $\mathbf{p}$.

4.2 *Monotonicity.* $\Pi(\mathbf{p})$ is a non-decreasing function of output prices and non-increasing function of input prices.

4.3 *Continuity.* $\Pi(\mathbf{p})$ is continuous in $\mathbf{p}$.

4.4 *Convexity.* $\Pi(\mathbf{p})$ is a convex function in $\mathbf{p}$.

4.5 *Homogeneity.* $\Pi(\mathbf{p})$ is linear homogeneous in $\mathbf{p}$.

Again, empirical analysis normally assumes, in addition,

4.6 *Differentiability.* Most empirical applications assume $\Pi(\mathbf{p})$ is twice differentiable. As was the case with the cost function, the profit function possesses two important corresponding derivative properties

(a) $$\frac{\partial \Pi(\mathbf{p})}{\partial p_i} = x_i, \qquad i = 1,\ldots,m+n \qquad \text{(Hotelling's Lemma)},$$

(b) $$\frac{\partial^2 \Pi}{\partial p_i \partial p_j} = \frac{\partial^2 \Pi}{\partial p_j \partial p_i} \Rightarrow \frac{\partial x_i}{\partial p_j} = \frac{\partial x_j}{\partial p_i} \qquad \text{(Symmetry)}.$$

4. A General Approach – Forms Linear-in-Parameters

4.1. *Parameterization of Economic Effects*

The main body of econometric and statistical technique requires models whose form is specified up to a finite vector of unknown parameters. This leads to the consideration of specific parametric production models which allow identification of particular economic effects, such as distribution and scale, while imposing no more maintained hypotheses than necessary on other aspects of technology. To a large extent we will be concerned with flexible representations of technologies, since flexibility is the issue which has led econometricians to seek alternatives to the first parametric production function, the Cobb–Douglas form [see Douglas and Cobb (1928)].

The objective of flexibility can be used to classify functional forms. Following Hanoch (1975a), we can specify the number of parameters required for representation of the economic effects discussed in Section 1. Consider an n input, one output production function $y = f(v_1,\ldots,v_n)$, with partial derivatives $f_i = \partial f / \partial v_i$ and $f_{ij} = \partial^2 f / \partial v_i \partial v_j$. Economic effects such as scale, distribution, and substitutability can in general be quantified in terms of the production function and its first and second derivatives. Consider the following classification of these effects:

Economic effect	Formula	Number of distinct effects
Output level	$y = f(\mathbf{v})$	1
Returns to scale	$\mu = \left(\sum_{i=1}^{n} v_i f_i\right)\Big/ f$	1
Distributive share	$s_i = v_i f_i \Big/ \sum_{j=1}^{n} v_j f_j$	$n - 1$
Own "price" elasticity	$\epsilon_i = v_i f_{ii} / f_i$	n
Elasticity of substitution	$\sigma_{ij} = \dfrac{-f_{ii}/f_i^2 + 2(f_{ij}/f_i f_j) - f_{jj}/f_j^2}{1/v_i f_i + 1/v_j f_j}$	$\dfrac{n(n-1)}{2}$

This table contains $(n+1)(n+2)/2$ distinct economic effects. These effects characterize the usual comparative statics properties of a production function at a point.[3] These formulae can be inverted to determine the function value and the first and second partial derivatives at a point in terms of economic effects,

$$
\begin{aligned}
f &= y, \\
f_i &= \mu y s_i / v_i, \\
f_{ii} &= \mu y s_i \epsilon_i / v_i^2, \\
f_{ij} &= [\sigma_{ij}(s_i + s_j) + \epsilon_i s_i + \epsilon_j s_j]\mu y / 2 v_i v_j, \qquad i \neq j.
\end{aligned}
$$

Hence, a necessary and sufficient condition for a functional form to reproduce comparative statics effects *at a point* without imposing restrictions across these effects is that it have $(n+1)(n+2)/2$ distinct parameters, such as would be provided by a Taylor's expansion to second-order.

[3] Exogenous technical change could be included by adding a variable t to the exogenous variables included in f. Then, $n+2$ economic effects would be added: the rate of technical change, $T = f_t/f$, the acceleration of technical change, $\dot{T} = (f_{tt}/f) - (f_t/f)^2$, and the rates of change of marginal products, $\dot{m}_i/m_i = f_{it}/f_i$. There would then be a total of $(n+2)(n+3)/2$ effects.

The development above in terms of a production function could equally well have been carried out in terms of a cost or profit function. Since the latter functions have $n+1$ arguments, compared to the n arguments of the production function, they may appear to permit a larger number of distinct effects involving first and second partial derivatives. However, the homogeneity properties of these functions reduce the number of independent parameters to $(n+1)(n+2)/2$, as before. For example, consider the cost function $C = C(y,\mathbf{r})$. Since C is homogeneous of degree one in $\mathbf{r}$, Euler's Theorem implies

$$\sum_{i=1}^{n} r_i C_i(y,\mathbf{r}) = C(y,\mathbf{r}),$$

$$\sum_{i=1}^{n} r_i C_{ij}(y,\mathbf{r}) = 0, \qquad j = 1,\ldots,n,$$

and

$$\sum_{i=1}^{n} r_i C_{yi}(y,\mathbf{r}) = C_y(y,\mathbf{r}),$$

where $C_i = \partial C/\partial r_i$, $C_{ij} = \partial^2 C/\partial r_i \partial r_j$, $C_y = \partial C/\partial y$, and $C_{yi} = \partial^2 C/\partial y \partial r_i$. These provide $n+2$ restrictions, known as the adding-up condition, the Cournot aggregation conditions, and the Engel aggregation condition, respectively. The number of distinct derivative conditions is therefore $(n+2)(n+3)/2 - (n+2) = (n+1)(n+2)/2$, as in the case of the production function.[4]

4.2. *Linear-in-Parameters Approximations*

Most of the flexible functional forms developed in the econometric literature can be viewed as linear-in-parameters expansions which approximate an arbitrary function. In general, such an expansion can be written in the form

$$f^*(\mathbf{x}) \approx f(\mathbf{x}) \equiv \sum_{i=1}^{N} a_i h^i(\mathbf{x}), \tag{1}$$

where f^* is the true function, f is the approximating functional form, the

[4] This argument can be applied to an n-input linear homogeneous production function to show that it has $n(n+1)/2$ distinct economic effects.

a_i are parameters, the h^i are known functions, and $\mathbf{x}$ is a vector of independent variables. In production applications, $\mathbf{x}$ may be input quantities or prices, or transformations of these variables (e.g., a log transformation). If $N = (n+1)(n+2)/2$ and a determinental condition (a non-singular Wronskian)[5] is satisfied at a point $\mathbf{x}^*$, then parameter values a_i can be found for which this expansion approximates the value of $f(\mathbf{x})$ and its first and second partial derivatives in a neighborhood of $\mathbf{x}^*$. We term an expansion with this property a *parsimonious flexible* form.

A common method of generating parsimonious flexible forms is by use of a Taylor's series expansion to second-order about a point $\mathbf{x}^*$. In this case, the known functions and corresponding parameters have the values

$$\begin{aligned}
&h^0(\mathbf{x}) = 1, && a_0 = f^*(\mathbf{x}^*), && \\
&h^i(\mathbf{x}) = x_i - x_i^*, && a_i = f_i^*(\mathbf{x}^*), && i = 1,\ldots,n, \\
&h^{ij}(\mathbf{x}) = (1/2)(x_i - x_i^*)(x_j - x_j^*), && a_{ij} = f_{ij}^*(\mathbf{x}^*), && i,j = 1,\ldots,n.
\end{aligned}$$

[For notational simplicity, the second-order terms in (1) have been reindexed in terms of i and j.]

A problem which arises when we consider parsimonious flexible functional forms as approximations to true functions is the accuracy of the approximation. If a flexible form is calibrated to provide a second-order approximation at a point, then the approximation is of this order only in a small neighborhood of this point. In other regions of interest, the form may be a poor approximation to the true function, and may even fail to satisfy basic properties of the true function such as mono-

[5] The Wronskian is the determinant

$$\begin{vmatrix}
h^0(\mathbf{x}^*) & h^1(\mathbf{x}^*) & h^N(\mathbf{x}^*) \\
\partial h^0(\mathbf{x}^*)/\partial x_1 & \partial h^1(\mathbf{x}^*)/\partial x_1 & \partial h^N(\mathbf{x}^*)/\partial x_1 \\
\vdots & \vdots & \vdots \\
\partial h^0(\mathbf{x}^*)/\partial x_n & \vdots & \partial h^N(\mathbf{x}^*)/\partial x_n \\
\partial^2 h^0(\mathbf{x}^*)/\partial x_1^2 & \vdots & \partial^2 h^N(\mathbf{x}^*)/\partial x_1^2 \\
\partial^2 h^0(\mathbf{x}^*)/\partial x_1 \partial x_2 & \vdots & \partial^2 h^N(\mathbf{x}^*)/\partial x_1 \partial x_2 \\
\vdots & \vdots & \vdots \\
\partial^2 h^0(\mathbf{x}^*)/\partial x_n^2 & \cdot & \partial^2 h^N(\mathbf{x}^*)/\partial x_n^2
\end{vmatrix}$$

When this determinant is non-zero, the coefficients a_i in (1) can be chosen so that the approximation to f has first- and second-order derivatives at $\mathbf{x}^*$ equal to those of f at $\mathbf{x}^*$.

tonicity or convexity.[6] Further, the qualitative implications of the calibrated approximation may depend on the point of approximation; this is true, for example, of separability, which involves properties of the true function beyond second-order (see Section 6). The economic effects of interest in comparative statics, while unrestricted at the point of approximation, can be strongly and perhaps implausibly related at different points in the domain of the expansion.

If a parsimonious flexible form is fitted to observations over an extensive domain, as is normally the case in econometric production analysis, then the fitted form will not in general be a second-order approximation to the true function at any chosen point. As a result, the comparative statics effects deduced from the approximation will bear a complex and perhaps misleading relationship to the corresponding effects for the true function. In particular, multivariate fits to the approximate function and its derivatives may fail to satisfy restrictions on parameters across equations, even when the true function satisfies properties implying these restrictions. This could lead the analyst to conclude incorrectly that the true function fails to satisfy the properties in question. For example, tests of "profit-maximizing behavior" based on symmetry restrictions across equations may be rejected in the system of fitted functions even if the property holds in the true system. Note that this conclusion depends critically on the assumption that the expansion is being fitted to data over a large domain; a second-order approximation at a point will satisfy symmetry restrictions across equations when the true system does.

A simple example may help to clarify the issues raised in the preceding paragraphs. Suppose a true one-input production function is $y = e^v$, exhibiting increasing returns at an increasing rate (for $v > 1$), $\mu = v$, and a positive own-price elasticity, $\epsilon = v$. Suppose we approximate this production function with an expansion in logarithms, $\log y = a_1 + a_2 \log v + a_3(\log v)^2$. The estimated returns to scale and own-price elasticity from the expansion are $\hat{\mu} = a_2 + 2a_3 \log v$ and $\hat{\epsilon} = (2a_3/\hat{\mu}) + \hat{\mu} - 1$, respectively. Suppose $\log v$ is normally distributed with mean $\log m$ and variance σ^2. Then, an ordinary least-squares fit of the parameters in the expansion converges in probability to $a_1 =$

[6] Some expansions, such as the Translog function discussed below, can never except in trivial cases satisfy monotonicity or convexity conditions over the entire positive orthant. Hence, it is important in using these expansions to test for the satisfaction of maintained hypotheses in regions of interest. In Appendix A.4 of this volume, Lau provides computational methods for verifying convexity.

$me^{\sigma^2/2}[1 - \log m + \frac{1}{2}(\log m)^2 - \sigma^2)]$, $a_2 = me^{\sigma^2/2}(1 - \log m)$, and $a_3 = (m/2)e^{\sigma^2/2}$. Alternatively, a second-order approximation to the true function at a point $v = m$ satisfies these formulae with $\sigma^2 = 0$.) Let $\hat{y}$, $\hat{\mu}$, and $\hat{\epsilon}$ denote the economic effects measured from the fitted expansion. Then, for example,

$$\frac{\hat{\mu}}{\mu} = e^{\sigma^2/2}\left(1 + \log\frac{v}{m}\right)\bigg/\frac{v}{m},$$

which attains a maximum of $e^{\sigma^2/2}$ at $v = m$. Hence, a second-order expansion at a point will underestimate the returns to scale effect except at the point. A fit to data for which $\log v$ is normal with mean $\log m$ and variance σ^2 yields an overestimate of returns to scale at the data mean.

Table 1 indicates the accuracy of the approximation to y, μ, and ϵ for three alternative expansions. In each case, the approximation is good (say, within 10 percent) only in a narrow range, and is particularly poor for small v where the expansions fail to satisfy monotonicity. The effect of fitting the expansion to log normal data with $m = 10$, $\sigma^2 = 1$ is a

TABLE 1

	Second-order fit $m = 1$			Second-order fit $m = 10$			Data fit when $\log v$ has mean $\log m = \log 10$, var $\sigma^2 = 1$		
v	$\frac{\log \hat{y}}{\log y}$	$\frac{\hat{\mu}}{\mu}$	$\frac{\hat{\epsilon}}{\epsilon}$	$\frac{\log \hat{y}}{\log y}$	$\frac{\hat{\mu}}{\mu}$	$\frac{\hat{\epsilon}}{\epsilon}$	$\frac{\log \hat{y}}{\log y}$	$\frac{\hat{\mu}}{\mu}$	$\frac{\hat{\epsilon}}{\epsilon}$
0.3	1.74	−0.68	−20.36	121.38	−83.55	−88.22	172.64	−137.75	−142.42
0.4	1.26	0.21	27.57	74.04	−55.47	−59.10	101.47	−91.46	−95.08
0.5	1.09	0.61	5.13	49.83	−39.91	−42.92	65.67	−65.81	−68.81
0.9	1.00	0.99	1.12	16.57	−15.64	−17.54	18.16	−25.79	−27.69
1.0	1.00	1.00	1.00	13.48	−13.03	−14.79	13.99	−21.48	−23.24
1.1	1.00	1.00	0.92	11.17	−10.98	−12.64	10.92	−18.10	−19.76
2.0	0.97	0.85	0.64	3.43	−3.05	−4.37	1.53	−5.02	−6.34
3.0	0.90	0.70	0.53	1.74	−0.68	−2.65	0.11	−1.12	−3.09
4.0	0.84	0.60	0.45	1.26	0.21	2.95	0.01	0.35	3.08
5.0	0.78	0.52	0.40	1.09	0.61	1.07	0.16	1.01	1.46
9.0	0.62	0.36	0.28	1.00	0.99	1.01	0.73	1.64	1.65
10.0	0.60	0.33	0.26	1.00	1.00	1.00	0.82	1.65	1.65
11.0	0.57	0.31	0.24	1.00	1.00	0.99	0.90	1.64	1.63
20.0	0.42	0.20	0.16	0.97	0.85	0.83	1.18	1.40	1.38
30.0	0.34	0.15	0.12	0.90	0.70	0.68	1.21	1.15	1.14
40.0	0.29	0.12	0.10	0.84	0.60	0.58	1.17	0.98	0.97
50.0	0.25	0.10	0.08	0.78	0.52	0.51	1.12	0.86	0.85
90.0	0.17	0.06	0.05	0.62	0.36	0.35	0.94	0.59	0.58

substantial overestimate of μ and ϵ in the range $3.68 \leqq v \leqq 27.18$, which contains 68 percent of the data. This example suggests that fitted expansions can be relatively non-robust with respect to the point of approximation or range of data available, and that considerable caution should be used in utilizing the models for extrapolative prediction or the testing of basic hypotheses on production structure.

In principle, the difficulty in obtaining accurate approximations in the large can be overcome by introducing additional parameters. On a closed bounded domain, the Bernstein–Weierstrauss approximation theorem shows that a continuous function can be approximated uniformly by polynomials.[7] In practice, the number of parameters required in these theorems to guarantee a specified level of accuracy is too large for empirical purposes. A theory of approximation in the large for production functions which incorporates the qualitative properties of the true functions such as monotonicity and convexity might produce tighter bounds on the number of parameters required; however, this topic is beyond the scope of this survey.

[7] Suppose the domain of interest is defined – by translation, normalization, and extension if necessary – to be the closed bounded set $\mathbf{S} = \{(v_1,\ldots,v_n) \geqq 0 | v_1 + \cdots + v_n \leqq 1\}$. Consider the class of functions f which are uniformly Lipschitzian on $\mathbf{S}$ with constant M; i.e., $|f(\mathbf{v}) - f(\mathbf{v}')| \leqq M\|\mathbf{v} - \mathbf{v}'\|$. Define a multivariate Bernstein polynomial

$$B_N(\mathbf{v}) = \sum_{(k_1,\ldots,k_n)\in\mathbf{K}} f(k_1/N,\ldots,k_n/N) b_N(\mathbf{v};k_1,\ldots k_n),$$

where $\mathbf{K}$ is the set of integer vectors $(k_1,\ldots,k_n)$ with $(k_1/N,\ldots,k_n/N) \in \mathbf{S}$, and

$$b_N(\mathbf{v};k_1,\ldots,k_n) = \{N!/(k_1!\cdots k_n!(N - k_1 - \cdots - k_n)!)\} v_1^{k_1} \cdots v_n^{k_n} (1 - v_1 - \cdots - v_n)^{N-k_1-\cdots-k_n}.$$

Given $\epsilon > 0$, if $N \geqq nM^2/\epsilon^2$, then $|f(v) - B_N(v)| \leqq \epsilon$ uniformly on the cube. To establish this result, define $\mathbf{K}_1 = \{k \in \mathbf{K} | \, \|\mathbf{v} - \mathbf{k}/N\| \leqq (n/4N)^{1/2}\}$ and $\mathbf{K}_2 = \mathbf{K} \backslash \mathbf{K}_1$. Then

$$\begin{aligned} |f(\mathbf{v}) - B_N(\mathbf{v})| &\leqq \sum_{k\in K_1} \left| f(\mathbf{v}) - f\left(\frac{\mathbf{k}}{N}\right)\right| b_N(\mathbf{v};\mathbf{k}) + \sum_{k\in K_2} M \frac{\|\mathbf{v} - \mathbf{k}/N\|^2}{\|\mathbf{v} - \mathbf{k}/N\|} b_N(\mathbf{v};\mathbf{k}) \\ &\leqq (n/4N)^{1/2} M \sum_{k\in K} b_N(\mathbf{v};\mathbf{k}) + M(n/4N)^{-1/2} \sum_{k\in K} \|\mathbf{v} - \mathbf{k}/N\|^2 b_N(\mathbf{v};\mathbf{k}) \\ &\leqq M(n/4N)^{1/2} + M(n/4N)^{-1/2} n \sum_{i=1}^{n} \frac{v_i(1 - v_i)}{N} \\ &\leqq M\left(\frac{n}{N}\right)^{1/2} \leqq \epsilon, \end{aligned}$$

where the second and third inequalities follow from properties of the multinomial distribution.

4.3. Common Linear-in-Parameters Forms

Table 2 provides, in summary form, a list of the most commonly used linear-in-parameters functional forms and their approximation characteristics. The historic Cobb–Douglas function, while not originally proposed as an approximation, can be viewed as a first-order expansion in $\log v_i$ about $v_i = 1$. This form allows free assignment of the output level, returns to scale, and distributive shares effects at a point of approximation, but allows no flexibility with respect to the substitution and own-price elasticity effects. The CES function adds one substitution parameter to the (linear homogeneous) Cobb–Douglas case. We have included this functional form in the table, even though it is not linear-in-parameters unless the substitution parameter ρ is known, because it is the basis for several linear-in-parameter expansions.

The concept of linear-in-parameters functional forms and the property of second-order approximation at a point are due to Diewert (1971), who introduced the generalized linear and generalized Leontief systems. This development was followed by the introduction of the translog functional form by Christensen, Jorgenson, and Lau (1971). A direct generalization of the Cobb–Douglas function, the translog form has been widely used as a framework for analysis of structural properties of production.

All the forms in Table 2 with the exception of the Quadratic have restrictions implying linear homogeneity, and under this restriction have $n(n+1)/2$ parameters, as required for a parsimonious flexible linear homogeneous function. In the absence of homogeneity restrictions, the forms having $(n+1)(n+2)/2$ parameters are Generalized Leontief, Translog, and Quadratic. With the exception of Generalized Cobb–Douglas and Generalized Concave forms, the functions in Table 2 can be interpreted as Taylor's expansions about a point. In this interpretation, first proposed explicitly by Lau (1974), Cobb–Douglas is a first-order expansion of $\log y$ in powers of $\log x_i$, and Translog is a second-order expansion. CES is a first-order expansion of y^ρ in powers of x_i^ρ. Generalized Leontief and Quadratic are second-order expansions of y in powers of $\sqrt{x_i}$ and x_i, respectively.

Generally, the forms in Table 2, or analogous forms that could be obtained using other series of functions as a basis for expansions, will provide equally satisfactory representations of an arbitrary production function at a point. Choice between them should be based on their quality as approximations to the true functions over the full domain of

TABLE 2
Common linear-in-parameters functional forms.[a]

Functional form	Formula	Restrictions
Cobb–Douglas [Douglas–Cobb (1928)]	$\log y = a_0 + \sum_{i=1}^{n} a_i \log x_i$	$\sum_{i=1}^{n} a_i = 1$ for linear homogeneity
CES [Arrow et al. (1961)][b]	$y^\rho = a_0 + \sum_{i=1}^{n} a_i x_i^\rho$	$a_0 = 0$ for linear homogeneity
Generalized Leontief/Linear [Diewert (1971)]	$y = a_0 + \sum_{i=1}^{n} a_i \sqrt{x_i} + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \sqrt{(x_i x_j)}$	$a_i = 0$, $i = 0,\ldots,n$ for linear homogeneity
Translog [Christensen–Jorgenson–Lau (1971)]	$\log y = a_0 + \sum_{i=1}^{n} a_i \log x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} (\log x_i)(\log x_j)$	$\sum_{i=1}^{n} a_i = 1$ and $\sum a_{ij} = 0$ for linear homogeneity
Generalized Cobb–Douglas [Diewert (1973b)]	$\log y = a_0 + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \log(x_i + x_j)/2$	$\sum_i \sum_j a_{ij} = 1$ for linear homogeneity
Quadratic [Lau (1974)]	$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$	
Generalized Concave [McFadden (Chapter II.2)]	$y = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i \phi^{ij}(x_i/x_j) a_{ij}$	ϕ^{ij} a known concave[c] function

[a] x is a vector of inputs or prices, depending on whether a direct or dual form is being considered. y is output, cost, or profit. Except for the case of self-dual forms [see Hanoch (Chapter I.2)], these formulae defined in terms of prices yield different technologies than are obtained from these formulae defined in terms of quantities. Linear homogeneous functions can also be represented by these forms by expressing all variables in relative terms.

[b] If

$$a_0 = 0 \quad \text{and} \quad \sum_{i=1}^{n} a_i = 1$$

in the CES formula, then a first-order expansion in ρ,

$$\log y = \sum_{i=1}^{n} a_i \log x_i + \frac{\rho}{2}\left[\sum_{i=1}^{n} a_i (\log x_i)^2 - \left(\sum_{i=1}^{n} a_i \log x_i\right)^2\right],$$

provides a linear-in-parameters translog form with second-order terms

$$a_{ii} = \frac{\rho}{2} a_i (1 - a_i) \quad \text{and} \quad a_{ij} = \frac{\rho}{2} a_i a_j, \quad i \neq j.$$

This expansion was first suggested by Kmenta (1967), and has been utilized by Sargan (1971) and Griliches and Ringstad (1971). An alternative derivation of this expansion from the translog function, based on the imposition of a common AES between input pairs, has been given by Denny and Fuss (1977).

[c] For a profit function, the ϕ^{ij} are convex functions.

interest, to the extent that this can be assessed *a priori*, and on the ease with which hypotheses of interest can be stated as restrictions on parameters.

5. Special Non-Linear Forms

5.1. *Elasticities of Substitution*

The flexible forms discussed in Section 4 can be viewed as extensions of simple functional forms where the extensions are constrained to remain linear-in-parameters. For example, Diewert's Generalized Leontief cost function is just such an extension of the cost function dual to a Leontief fixed coefficient production function. The Translog extends the Cobb–Douglas function and the Quadratic extends a linear function under the same linear-in-parameters constraint. While linearity is retained, it is at the cost of introducing a large number of parameters into the analysis. The variants of simple functional forms surveyed in this section are characterized by non-linearity in parameters. This fact makes them less useful in general for econometric estimation than those forms surveyed in Section 4. However, in some cases, non-linearity is compensated for by parsimony in parameters. An example discussed in this section is Hanoch's CRESH–CDE form for use in the study of factor substitution.

Most of the forms surveyed in this section were devised to generalize, using as a few additional parameters as possible, two restrictive features of the maintained hypotheses concerning substitution effects of the original ACMS function. First, in the two-factor case, the elasticity of substitution is constrained to be constant, and there is no apparent technological justification for this restriction. Second, extension of the CES function to more than two factors requires, with unimportant exceptions, the imposition of the maintained hypothesis that all partial AES are equal and constant [Uzawa (1962)]. In the multiple factor case, it is not clear that the AES will be the desired concept of the elasticity of substitution (ES). The attempts to apply this concept to the case of more than two inputs have produced various definitions [McFadden (1963)]. As indicated by Mundlak (1968b), those definitions differ in two major respects: (1) the variables which are held constant in the underlying economic experiment and (2) the number of variables which are involved in the operation. If we denote $\hat{v} = \mathrm{d} \ln v$, and assume that all derivatives are evaluated at an equilibrium point, we can distinguish between one-factor one-price ES (OOES), $\hat{v}_j/\hat{r}_i$ (the AES is of this form), two-factor one-price ES (TOES), $(\hat{v}_j - \hat{v}_i)/\hat{r}_i$, and two-factor two-price ES (TTES), $(\hat{v}_j - \hat{v}_i)/(\hat{r}_i - \hat{r}_j)$. The last is the "usual" definition of ES. Each of these concepts can be evaluated at constant output, cost, or marginal cost. Each of these alternatives corresponds to a different factor demand curve, where the prices not involved in the operation are held constant. However, it is also possible to hold constant the quantities of the factors which are not involved in the operation. In one extreme "short-run" case we have the direct ES (DES), which is a TTES with all factors other than those involved in the operation held constant. In the extreme "long-run" case, we have the shadow ES (SES), in which all quantities are allowed to vary. We can also have mixed situations in which the quantities of some factors and the prices of other factors are held constant. Detailed discussions of various definitions of the ES are given in McFadden (1963), Hanoch (Chapter II.3), and Lau (Chapter I.3). All these forms collapse to a common definition in a two-factor linear homogeneous production function. This is due to the singularity of the Hessian matrix, and therefore it cannot be used as an indication that any of the above expressions is a generalization of the two-factor measure [Mundlak (1968b, p. 231)].

In summary, once we depart from the two-input case we confront the following problems in attempting to develop production functions from the point of view of the elasticity of substitution:

(a) There is no unique natural generalization of the two factor definition of the ES. The different definitions involve different combinations of the elements of the underlying Hessian matrix. It is therefore reasonable to deal with the Hessian elements directly. The AES comes close to this approach. Other than that, it has no particular advantage over the others and perhaps the reference to it as an elasticity of substitution is misleading. It is simply proportional to the cross elasticity in the constant output factor demand function. We conclude that the selection of a particular definition should depend on the question asked.

(b) The choice of an ES does *not* imply constancy of the elasticity; this is an added hypothesis which may not hold in reality. As a result, there is no direct relationship between the concept of the ES to be used and the algebraic form of the production function.

Non-linear forms have been analyzed primarily in terms of the AES, and in the pages which follow we will maintain the classification in terms of the AES. However, we note that for the reasons above, it might also be useful to pursue a classification similar to that in Section 4.

5.2. *Variants of the Cobb–Douglas Function*

(1) *Variable elasticity of substitution (VES) production function* [Revankar (1971)]. This function was devised to relax the assumption of constant AES in the two-factor case. It takes the form

$$y = \alpha_0 v_1^{\alpha_1}(v_2 + \gamma_1 v_1)^{\alpha_2},$$

and has an AES $= 1 + \beta(v_1/v_2)$ where β is a function of the production function parameters. It is considered a variant of the CD form since the AES varies around one for $\beta \neq 0$ and variations in relative inputs.

(2) *Constant marginal share (CMS) function* [*Bruno* (1968)]. This function is explicitly a generalization of the CD form. It can be expressed as

$$y = \alpha_0 v_1^{\alpha_1} v_2^{\alpha_2} - \gamma v_2,$$

and has an AES $= 1 - (\gamma\alpha_1/\alpha_2)(v_2/y)$.

(3) *Transcendental production function* [Halter, Carter, and Hocking (1957)]. This function has the form

$$y = \alpha_0 v_1^{1-\alpha} v_2^{\alpha} e^{\gamma_1 v_1 + \gamma_2 v_2},$$

and has an AES $= (1 - \alpha + \gamma_1 v_1)(\alpha + \gamma_2 v_2)/((1-\alpha)(\alpha + \gamma_2 v_2)^2 + \alpha(1 - \alpha + \gamma_1 v_1)^2)$ which reduces to unity when $\gamma_1 = \gamma_2 = 0$.

5.3. *Variants of the CES Function*

Most of the variants of the CES function can be seen as the result of attempting to eliminate the assumption contained in the multifactor CES formulation, namely, equality of all partial AES [Uzawa (1962), McFadden (1963)]. One extension which relaxes this restriction is the nested CES function [Sato (1967), see also McFadden (Chapter IV.1)]. This form has not been used extensively in empirical work due to its complex nature when extended to more than three factors [however, see Mundlak and Razin (1969, 1971)]. Another avenue for CES-like extensions is the class of implicitly additive forms introduced by Hanoch (1975a). The direct form is

$$F(\mathbf{v},y) = \sum_{i=1}^{n} F^i(v_i,y) \equiv 1, \qquad (3)$$

where F^i are functions with properties imposed to insure that the implied explicit production function satisfies the maintained hypotheses of Sections 3.1 and 3.2. The dual, or indirect form, is

$$G(\mathbf{r}/c,y) = \sum_i G^i(r_i/c,y) \equiv 1,^8 \tag{4}$$

where G^i are functions with properties imposed to insure that the implied cost function c satisfies the maintained hypotheses of Section 3.3. This class of functions contains as special cases the direct forms which have constant ratios of AES; such as the one derived by Mukerji (1963) and Gorman (1965) [and used by Dhrymes and Kurz (1964)] and the CRESH form developed by Hanoch (1971). The indirect form contains as special cases functions which exhibit constant differences in AES such as the CDE form of Hanoch (1971). The Mukerji form uses the transformation $F^i(v_i,y) = D_i(v_i^{d_i}/y^h)$, while the CRESH form uses the transformation $F^i(v_i,y) = D_i(v_i/y^h)^{d_i}$. The CDE form uses the transformation $G^i(r_i/c,y) = D_i(y^h r_i/c)^{d_i}$. In these transformations, D_i, d_i, and h are parameters. Detailed discussion of these transformations and extensions can be found in Hanoch (1975a).

Estimating equations for the indirectly additive class contain a small number of parameters when compared with the general linear expansions of Section 4. For example, consider the CRES (constant ratio elasticity of substitution – non-homothetic) form introduced by Hanoch (1975a),

$$\sum_i D_i y^{-e_i d_i} v_i^{d_i} \equiv 1, \tag{5}$$

where D_i, d_i, and e_i are parameters. Using the first-order conditions for profit maximization, one obtains the set of equations

$$\log v_i = A_i - a_i \log(r_i/r_1) + h_i \log y + (a_i/a_1)\log v_1, \qquad i = 1,\ldots,n,$$

where

$$a_i = 1/(1-d_i), \quad A_i = \log(D_i d_i/D_1 d_1)^{a_i}, \quad h_i = a_i(e_1 d_1 - e_i d_i),$$

and

$$\text{AES}_{i,k}/\text{AES}_{j,k} = a_i/a_j.$$

[8] For extensions of the indirectly additive class to the multiple output case see, Hanoch (Chapter II.3).

The above set of equations is non-linear in a_i and simultaneous in v_i but could be estimated using non-linear simultaneous equations procedures currently available. Note that there are $3n-2$ parameters to estimate in the system of equations (5) as compared with $(n+2)(n+1)/2$ for the second order linear-in-parameters approximations. The cost (in addition to the non-linearity) is the maintained hypothesis of implicit separability which is reflected in the fact that only two input prices appear as exogeneous variables in each demand equation. This is an example of the importance of separability assumptions for functional specification. It is to this issue that we now turn.

6. Separability: Functional Implications and Tests

6.1. *Basic Concepts*

Separability has various implications. It allows decentralization in decision-making or equivalently, optimization by stages. This opens up the possibility of consistent multi-stage estimation which may be the only feasible procedure when large numbers of inputs and outputs are involved; specifically, when applying the relatively simple concept of a production function to complex organizations. Historically, separability has played an important role in the specification of functional forms. The Cobb–Douglas and CES functions are explicitly strongly separable. Hanoch's (1971) CRESH–CDE class of functions is implicitly strongly separable. Sato's (1967) nested CES specification is strongly separable with respect to the highest level partition and then strongly separable within each sub-aggregate.

To define separability, we first denote the set of n inputs by $\mathrm{N}=\{1,...,n\}$. A partition S of N is given by $\{\mathrm{N}_1,...,\mathrm{N}_S\}$ where $\mathrm{N}=\mathrm{N}_1\cup\mathrm{N}_2\cdots\cup\mathrm{N}_S$, and $\mathrm{N}_r\cap\mathrm{N}_t$ is empty for $r\neq t$. Separability is characterized by the independence of the marginal rate of substitution between a pair of inputs from changes in the level of another input, i.e.,

$$\frac{\partial(f_i/f_j)}{\partial v_k}=0, \tag{6}$$

or $f_jf_{ik}-f_if_{jk}=0$. We say that f is *strongly separable* (SS) with respect to the partition S if (6) holds for all $i\in\mathrm{N}_r$, $j\in\mathrm{N}_t$, and $k\notin\mathrm{N}_r\cup\mathrm{N}_t$. The function is *weakly separable* (WS) with respect to the partition **S** if (6) holds for all $i,j \in \mathbf{N}_r$, and $k \notin \mathbf{N}_r$. Note that these properties may hold at a point or globally.

Goldman and Uzawa (1964) showed that a function $f(\mathbf{x})$ is globally SS with respect to the partition **S** ($S > 2$) if and only if $f(\mathbf{x}) = F\{\sum_{t=1}^{S} f^t(\mathbf{x}^t)\}$ where F is monotone increasing and $f^t(\mathbf{x}^t)$ is some function of $\mathbf{x}^t$. The function is globally WS with respect to the partition **S** if and only if it is of the form

$$f(\mathbf{x}) = G\{g^1(\mathbf{x}^1),\dots,g^S(\mathbf{x}^S)\}. \tag{7}$$

Berndt and Christensen (1973b) related separability to AES and obtained the result that any strictly quasi-concave homothetic production function $f(\mathbf{v})$ is WS with respect to the partition S *at a point* if and only if $\text{AES}_{ik} = \text{AES}_{jk}$ at that point for all $i,j \in \mathbf{N}_r$, $k \notin \mathbf{N}_r$. Similarly, the function is SS at a point if and only if $\text{AES}_{ik} = \text{AES}_{jk}$ for all $i \in \mathbf{N}_r$, $j \in \mathbf{N}_t$, $k \notin \mathbf{N}_r \cup \mathbf{N}_t$. Furthermore, if $n = S$, then all AES_{ik}, $i \neq k$, are equal. If this function is globally SS for any input combination then $f(\mathbf{v}) = F(\sum_{i=1}^{n} \alpha_i v_i^{\rho})$, a homothetic transformation of a CES function.

Finally, Berndt and Christensen showed that if $f(\mathbf{v})$ is homothetically separable then the dual cost function $C(y,\mathbf{r})$ is weakly separable so that

$$C_j C_{ik} - C_i C_{jk} = 0 \tag{8}$$

holds as well as (6).

In proving these theorems Berndt and Christensen use a result obtained by Lau (Chapter I.3) to the effect that the cost function is WS(SS) with respect to the partition **S** in input prices if and only if $f(\mathbf{v})$ is homothetically WS(SS) with respect to the partition **S** in input quantities.

The role of homogeneity of f in the Berndt–Christensen results is analyzed by Russell (1975), who extends the results to the case of non-homothetic production functions.

Separability results comparable to those obtained by Berndt and Christensen are developed in terms of cost and profit functions by McFadden (Chapters I.1 and IV.1) and Lau (Chapter I.3).

One important application of separability is in the derivation of value-added functions. If the gross output production function is weakly separable in primary inputs then a net output or value-added function can be defined and used for analysis. This issue is pursued by Bruno (Chapter III.1) and Denny and May (Chapter III.3).

6.2. *Separability in Forms Linear-in-Parameters*

Since the separability constraints (6) and (8) depend on second-order partial derivatives, functional forms linear-in-parameters must be at least of the second-order in the variables to contain separability as a testable implication. For example, the Cobb–Douglas function, which is of the first-order in logarithms, maintains separability since $(\partial^2 \log f)/(\partial \log v_i)(\partial \log v_j) = 0$ for all i,j, thus satisfying (6). The class of second-order approximation functions then will be the linear in parameters class necessary in general to test separability. Separability tests of production structures using the translog specification can be found in Berndt and Christensen (1973a, 1974), Berndt and Wood (1975), Denny and Fuss (1977), and Denny and May (Chapter III.3). Similar testing of the structure of utility functions appears in Christensen, Jorgenson, and Lau (1975) and Jorgenson and Lau (1975a). An alternative approach to testing separability, in the framework of the multistage Sato function, appears in Mundlak and Razin (1971).

The above tests fall into two categories. The first category is that of "exact" tests. These tests result from the imposition of the null hypothesis of separability for all possible values of the exogenous variables. The second category consists of "approximate" tests, where the null hypothesis is imposed only at a point of approximation, utilizing the notion of the function as a second-order Taylor series expansion. Berndt–Christensen and Berndt–Wood use the exact tests, Denny–Fuss and Denny–May use the approximate ones, while Christensen et al., and Jorgenson and Lau use both (under the terminology "intrinsic" and "explicit"). Exact tests would seem to be preferable if no additional constraints are imposed, since a single reject/non-reject decision is globally applicable. Unfortunately, with second-order expansions this is not the case. Blackorby et al. (1977a) and Denny–Fuss (1977) have shown that the restriction of global weak separability implies either strong separability within the partitioned sub-aggregates, or strong separability between aggregates. For example, suppose G in (7) is translog. Then either each $g^i(\mathbf{x}^i)$ is Cobb–Douglas in $\mathbf{x}^i$ or G is Cobb–Douglas in g^i. These results can also be found in Jorgenson and Lau (1975) for the case of utility functions. We are left with a tradeoff between tests which impose extraneous restrictions and those which depend on the data point chosen as the point of approximation. While the issue remains unresolved, one possible procedure is to explore higher-order expansions [Lau (1977)], which unfortunately requires the introduction of a large number of additional parameters. Another approach is to explore forms non-linear-in-parameters, to which we now turn.

6.3. *Separability in Forms Non-Linear-in-Parameters*

We begin by illustrating a procedure suggested by Mundlak (1973a) for generating non-separable functions which contain less than the $(n+1)\times(n+2)/2$ independent parameters of the second-order approximations. To sketch the approach to this problem, let

$$y = f(\mathbf{v}) = (g * h)(\mathbf{v}), \tag{9}$$

where $f(\mathbf{v})$ is the production function, g and h are two arbitrary functions, and $*$ is an arbitrary operator; i.e., addition, multiplication, exponentiation, or composition.[9] It can be shown that g and h can both be separable while $f(\mathbf{v})$ itself is not separable.

To illustrate the use of this approach, let $*$ be addition so that (9) becomes

$$f(\mathbf{v}) = g(\mathbf{v}) + h(\mathbf{v}). \tag{10}$$

Then to evaluate (6) we can write

$$\begin{aligned} f_if_{jk} - f_jf_{ik} &= (g_i + h_i)(g_{jk} + h_{jk}) - (g_j + h_j)(g_{ik} + h_{ik}) \\ &= (h_ih_{jk} - h_jh_{ik}) + (g_ig_{jk} - g_jg_{ik}) \\ &\quad + (h_ig_{jk} - h_jg_{ik}) + (g_ih_{jk} - g_jh_{ik}). \end{aligned} \tag{11}$$

We now note that h and g can both be separable so that the first two terms on the r.h.s. of (11) vanish. Furthermore, we can select one of the functions to be linear. For instance, let $g_i \neq 0$ for at least one i and $g_{ir} = 0$ for all i and r. Thus, if g is linear and h is separable we get

$$f_if_{jk} - f_jf_{ik} = g_ih_{jk} - g_jh_{ik}. \tag{12}$$

For f to be non-separable with respect to i and j it is sufficient that (12) differs from zero. For instance, we can assume h to be a CD with α_j being the output elasticity with respect to the jth factor, so that $h_{ik} = \alpha_i\alpha_k(h/v_iv_k)$. Then (7) becomes

$$g_ih_{jk} - g_jh_{ik} = \frac{\alpha_k}{v_k}\, h \left(g_i \frac{\alpha_j}{v_j} - g_j \frac{\alpha_i}{v_i}\right). \tag{13}$$

[9] By composition $g*h$, we shall mean that h becomes an argument of g; i.e., $(g*h)(\mathbf{v}) \equiv g(\mathbf{v}, h(\mathbf{v}))$.

(13) is equal to 0 for all i and j if and only if

$$\frac{g_j\alpha_i}{v_i} = \frac{g_i\alpha_j}{v_j}, \tag{14}$$

which is impossible except at a point.

Note that (14) could be used as an approximate test of separability at a point. In contrast to the translog function, the maintained hypotheses involve only $2n$ parameters.

We can further illustrate the above procedure by using it to generate a second-order approximation form which can be used to test separability among outputs within the class of exact tests.

Let $C(\mathbf{y},\mathbf{r})$ be a cost function dual to the distance function $f(\mathbf{y},\mathbf{v}) = 0$ where $\mathbf{y}$, $\mathbf{v}$, $\mathbf{r}$ are output, input, and input price vectors, respectively. Suppose

$$C(\mathbf{y},\mathbf{r}) = (g * h)(\mathbf{y},\mathbf{r}) \equiv g(\mathbf{y},h(\mathbf{r})), \tag{15}$$

where $*$ is a "composite function" operator and $h(\mathbf{r})$ is a vector consisting of elements $h_{ij}(\mathbf{r})$, $i,j = 1,\dots,n$. Let

$$h_{ij}(\mathbf{r}) = \sum_r \sum_s \alpha_{ij.kl}(r_k r_l)^{1/2} \quad \text{and} \quad g(\mathbf{y},h) = \sum_i \sum_j h_{ij}(y_i y_j)^{1/2}. \tag{16}$$

The resultant function is

$$C(\mathbf{y},\mathbf{r}) = \sum_i \sum_j \sum_k \sum_l \alpha_{ij.kl}(y_i y_j r_k r_l)^{1/2}, \tag{17}$$

which was analyzed by Hall (1973) and implemented empirically by Burgess (1976). Separability of the form $f^1(\mathbf{y}) = f^2(\mathbf{v})$ can be tested by imposing the restrictions $\alpha_{ij.kl} = a_{ij}a_{kl}$ [Hall (1973)] which results in an exact test. We arrived at the above form by combining two generalized Leontief specifications. Of course (17) still contains a large number of parameters, limiting its usefulness empirically. Nevertheless, this method of combining functions may prove useful for achieving a particular property with an efficient use of parameters.[10]

[10] The two-stage nested functional form developed in Fuss (1970) [see also Fuss (1977b) and Fuss and McFadden (Chapter II.4)] combines two generalized Leontief cost functions using a composite function rule much like that employed by Hall. This construction provides exact tests (in the sense used in the text) of the flexibility of the underlying technology.

7. Econometric Estimation of Production Parameters

The functional forms set out in Sections 4 and 5 characterize systematic relationships between economic variables, but take no account of the random effects which enter the determination of measurement of these variables. In application the stochastic specification is an intrinsic part of the specification of the production model. It should be emphasized that a specification of the model should be guided by the visualization of the *true* process and this is determined by nature and not by the econometrician. Hence the object is not to choose a specification that justifies a particular statistical procedure but on the contrary, to provide a general framework which allows for discrimination between various alternatives, as well as to examine the "robustness" of procedures dictated by the various alternatives.

Relations between *measured* production variables will in general contain stochastic components introduced at four levels: (1) the technology of the production unit, (2) the environment of each firm, particularly the market environment, (3) the behavior of the production unit, and (4) the process of observation, which often involves aggregation over commodities, production units, and time; direct errors in measurement; and incomplete observation. We discuss in turn each source of error.

Variations in technology from one production unit to the next may arise from specific or unit effects *known* to the production unit but not to the econometrician; such as management efficiency, availability and quality of specific factor inputs, and the presence of non-market inputs. They may also arise from effects which are *unknown* to the production unit at the time decisions are made. Examples are effects due to breakdown, weather, random variations in factor efficiency, and variations in quality control. The importance of the distinction between these two sources of variation [Mundlak and Hoch (1965)] is that effects known to the production unit enter the process of optimization and will be transmitted to the chosen input levels, whereas these chosen levels cannot depend on the realized values of random effects which are unknown at the time input decisions are made.[11] The statistical implications of this distinction are that observed factor inputs will be

[11] This is to some extent an oversimplification, because if production is performed by stages the error of one stage becomes a known error of higher stages [Mundlak (1963)].

endogenous if random effects are known to the production unit, and potentially exogenous if they are not.

The environment of a production unit includes a description of the markets in which input purchases and output sales must be made; the information available to the production unit at the time it makes decisions on market conditions; levels of non-market inputs; and the degree of organizational pressure or slack; as well as more general information on societal pressures on production unit decisions. For example, firms may face competitive input markets, and may purchase inputs of unknown quality in these markets at known prices, with the result that prices per efficiency unit of input are uncertain. Alternately, firms may find it necessary to contract for purchases or sales in some markets before other markets open, making relative prices uncertain. For instance, the purchase of durable inputs precedes the knowledge of all future prices of outputs and related inputs. If some markets are non-competitive, then stochastic components in demand or supply for non-competitive commodities will influence firm decisions and the resulting prevailing prices and quantities. In some cases it may be important to distinguish between stochastic effects on market equilibrium which are known to the firm, and thus part of its decision function, and those which are unknown to the firm. The knowledge need not be perfect for the argument to hold. The former will make observed prices endogenous; the latter makes them potentially exogenous.

Production unit behavior introduces stochastic components via deviations from idealized behavior patterns, as for example, failure in profit maximization to achieve exactly the desired marginal products of inputs. Such errors may arise from the finite computational ability of firms, from explicit calculations of computation costs versus expected gains, from satisficing behavior, or from firm objective functions which differ from those postulated in a maintained hypothesis by the econometrician. We note that some of these effects may introduce systematic biases into behavioral responses, and into the resulting observation. For example, Mundlak and Vulcani (1973) consider firm utility maximization, with utility depending not only on profit, but also on other variables like uncertainty and the leisure component in a production plan. It then follows that classical first-order conditions for profit maximization misspecify the true behavioral conditions, and therefore going from direct estimation of the production function to estimation of a system containing erroneous first-order conditions can be expected to worsen the quality of estimates [Mundlak (1973b)]. This caveat applies quite generally to the use of indirect forms or behavioral equations in estimating technological parameters; these forms require maintained hypotheses, such as profit maximization, in addition to those required by the basic specification of the technology. If these hypotheses prove to be false, then inferences on technology conditional on such hypotheses will be negated unless the estimation procedures can be shown to be robust. One such robust procedure, a direct estimation of the production function with prices serving as instrumental variables, is examined below. The robustness follows from the fact that even under a broader formulation, profit is considered to be an important argument in the utility function of the firm, and, ceteris paribus, prices have the same effect on quantities as in the neoclassical theory.

Broadening the framework of the analysis by allowing the utility function to include other variables in addition to profit leads to a duality relationship between technology and what may be referred to as a profit-like function – that is, a function which behaves like a profit function but whose arguments are some combination of actual prices and "prices" of the other variables that enter into the utility function. We can refer to the outcome of such combinations as pseudo prices. The profit-like function is the dual of the true production function. The use of profit rather than a profit-like function in empirical analysis can be considered as an approximation, the quality of which is to a large extent an empirical question. If however, it turns out that in a particular situation the approximation cannot be justified, the question is what information can be derived by working under the assumption that the system behaves *as if* the first-order conditions for profit maximization were met. Basically, this is a question of tracing the consequences of specification error in some equations on the model as a whole. If the technology is more stable than behavior, it may still be identified through the use of the first-order conditions for profit maximization. If on the other hand behavior is more stable, we derive behavioral equations which behave like reduced-form equations of a structure that is not fully identified.

In addition to stochastic components introduced in the technology and behavior of the production unit, there are observation errors introduced in the process of measurement of variables by the econometrician. First, classical measurement errors may occur in the process of soliciting, recording, and processing data. Second, a variety of sources of error, which can be lumped under the term aggregation errors, occur because of an inexact or ambiguous correspondence between ideal and practical definitions of variables. Further, "ideal" aggregation is determined by the true functional form, which is itself to be determined in the analysis. Thus, any given practical procedure of aggregation may lead to different kinds of aggregation error for alternative "true" production functions. Consequently, other things being equal, the aggregation error may influence the selection of a functional form in favor of the form for which the error is minimal. The aggregation problem occurs in various phases of the analysis. Aggregation over detailed commodity classifications (e.g., labor services distinguished by individual) to relatively homogeneous categories (e.g., labor services of stenographers) introduces errors. In the case of broad commodity classes, such as "capital" and "labor", these errors may be sufficiently major to influence the interpretation of the "technology". Aggregation over production units or through time may be dictated by the feasibility of data collection, or in the case of macroeconomic relationships, may be an objective of the analysis. Third, errors may arise because variables which are difficult or impossible to measure exactly are replaced by proxies, as for example the use of an average mortgage interest rate for a firm as a proxy for the actual interest rates on mortgages on specific structures.

In view of the complexity of the stochastic structure of production systems it should be clear that there is no simple universal estimation procedure. There are several alternatives whose merits depend on the relative strength of the various error components. In order to characterize these alternatives we note that the production function and the set of equations describing the first-order conditions for profit maximization constitute a complete system. The reduced-form of the system gives the product supply and factor demand equations. The profit function is an identity in the reduced-form equations.[12]

The main approaches to estimation are:

(1) direct estimation of the production function,
(2) estimation of the first-order equations,
(3) estimation of the reduced-form equations,
(4) estimation of the dual functions.

The selection of a particular approach depends not only on the stochastic specification but also on the functional form. However, in

[12] In this general discussion, we assume that all the variables are determined without any constraint on the maximization. Thus, the reduced form equations are long-run behavioral equations. If some variables are fixed, the reduced-form equations will include such constraints and thus result in short-run equations [Mundlak (1963)].

order to review the main points which have appeared in the literature dealing with the stochastic part of the model, we follow an example which assumes a very simple functional form – a Cobb–Douglas with one input only. We carefully specify and allow for the main sources of variations that have been discussed above and trace their effects on the various estimators considered. The discussion is oriented toward a cross-section analysis of firms. Some comments are also made on the possibilities which exist when there are repeated observations on firms. The specifications are listed as maintained hypotheses, and are not necessarily intended to represent an order of plausibility.

Example. An econometrician observes data on labor input (L), output (Y), and wage rate measured in output units (w) for a cross-section of firms, indexed $i = 1,\dots,T$. He wishes to estimate the elasticity of output with respect to labor input. The following maintained hypotheses are imposed.

7.1. *Technology*

7.1.1. *Variables*

Maintained Hypothesis 1. The technological possibilities of each firm are completely defined by two variables, the single variable input labor and a single output. There are no other variables such as capital, raw materials, knowledge, secondary outputs, etc., which vary systematically across the sample and enter the determination of technological possibilities.

7.1.2. *Functional Form*

Maintained Hypothesis 2. Each firm has the same technological possibilities, except for random effects due to (1) specific environment, management efficiency, and local labor quality, which will be referred to as the *firm effect*, and (2) breakdowns, weather, random variations in worker efficiency, which will be referred to as the *non-systematic error*. The technological possibilities have the Cobb–Douglas functional form

$$Y_* = AL_*^{\beta} e^{\epsilon+\lambda}, \tag{18}$$

where A and β are parameters, Y_* and L_* are the "true" values of output and labor input, ϵ is the firm effect and λ is the non-systematic error. These errors are normalized so that $E\epsilon = E\lambda = 0$.

[Note: Alternative specifications might be: (1) a production function other than Cobb–Douglas, or (2) firm-to-firm variation in parameters, such as β, or a more general variation in production possibilities across firms.]

7.2. *Environment*

7.2.1. *Market Structure*

Maintained Hypothesis 3. The firm faces competitive input and output markets. The relative price in these markets varies across firms, and is non-stochastic and fixed in repeated samples.

[Note: An alternative specification might be a non-competitive input or output market with relative prices endogenous and depending on firm behavior.]

7.2.2. *Information Available to the Firm*

Maintained Hypothesis 4. At the time the firm must choose its labor input, it knows the true production function, *except* for the non-systematic error λ about which the firm forms expectations. The firm measures its "true" input and output levels without error. In particular, the firm has no ambiguity about the "quality" of input or output. The *firm* measures the relative price of labor in terms of output with a random error, $\tilde{w} = w_* e^{\xi}$, where w_* is the "true" real wage, ξ is the error, and $\tilde{w}$ is the relative price of labor seen by the firm. The source of the random error ξ may be uncertainty about price at the time the relative price is measured; e.g., the firm may measure the money wage without error and forecast the output price with error, so that the ratio of the money wage to the output price, or real wage, is measured with error. As a first approximation it is convenient to assume that $E(\xi) = 0$. In the present context this is a very restrictive assumption for it indicates that the log forecast price is on the average equal to the log true price. Therefore, eventually we shall trace the consequences of the elimination of this assumption.

[Note: Alternative and supplementary specifications might be: (1) that the firm is uncertain about its true production function, (2) that the firm makes errors in measuring the amount of "true" labor in the labor quantity it observes because of an unknown "quality" factor, or (3) that the firm exhibits some systematic bias in measuring the relative price of labor.]

7.3. *Firm Behavior*

7.3.1. *Market Posture of the Firm*

Maintained Hypothesis 5. The firm attempts to maximize competitive profit, given the information available to it and the *point* expectation that $\lambda = 0$, by a choice of the labor input. The quantity sold and actual profit are determined by the actual value of λ.

Analysis: With the point expectation $\lambda = 0$, the firm "sees" the production function

$$Y_* = A\mathrm{e}^{\epsilon} L_*^{\beta}, \tag{19}$$

and relative input price $\tilde{w}$. It then "sees" the profit (measured in output units)

$$\pi = A\mathrm{e}^{\epsilon} L_*^{\beta} - \tilde{w} L_*. \tag{20}$$

The firm chooses L_* to maximize (20), setting the marginal product of labor equal to the real wage that it "sees",

$$\partial Y_* / \partial L_* = \beta A \mathrm{e}^{\epsilon} L_*^{\beta - 1} = \tilde{w}. \tag{21}$$

Errors in optimization can be subsumed in the random error ξ in forecasting the real wage. In this case, ξ may be subject to a firm effect, but this in turn makes the assumption of $E(\xi) = 0$ even more restrictive. As indicated, we return to this question later. From (21),

$$L_* = (\tilde{w} / \beta A \mathrm{e}^{\epsilon})^{-1/(1-\beta)}, \tag{22}$$

$$Y_* = A\mathrm{e}^{\epsilon + \lambda} L_*^{\beta} = (A\mathrm{e}^{\epsilon})^{1/(1-\beta)} \beta^{\beta/(1-\beta)} \tilde{w}^{-\beta/(1-\beta)} \mathrm{e}^{\lambda}, \tag{23}$$

where L_* is the "true" input, Y_* is the "true" output. The firm's "expected" output is given by (23) with $\lambda = 0$. The profit which the firm would receive from the "true" input–output combination if $\tilde{w}$ were the "true" relative price is

$$\tilde{\Pi} = Y_* - \tilde{w} L_* = \{A\mathrm{e}^{\epsilon} (\beta / \tilde{w})^{\beta}\}^{1/(1-\beta)} (\mathrm{e}^{\lambda} - \beta). \tag{24}$$

"Expected" profit for the firm is given by (24) with $\lambda = 0$. Finally, the profit the firm actually receives from the "true" input–output combination with the true relative price $w_* = \tilde{w}\mathrm{e}^{-\xi}$ is

$$\Pi_* = Y_* - w_* L_* = \{A\mathrm{e}^{\epsilon}(\beta/\tilde{w})^{\beta}\}^{1/(1-\beta)}(\mathrm{e}^{\lambda} - \beta \mathrm{e}^{-\xi}). \tag{25}$$

As a consequence of the forementioned hypotheses, true input, output, and profit satisfy (22), (23), and (25).

[Note: Alternative specifications of firm behavior are: (1) non-competitive behavior rules (even in the face of competitive markets), (2) objectives other than profit maximization (e.g., sales maximization, managerial tastes), (3) alternative models of expectation formation, particularly where the firm has some prior beliefs on the likelihood of various λ and ξ, and (4) treatment of risk aversion and a "utility" function of profits.]

7.4. *Observed Data*

7.4.1. *Relation of Observed and "True" Series*

Maintained Hypothesis 6. The econometrician observes the "true" relative wage, labor input, and output with error (but without systematic bias); specifically $w = w_* \mathrm{e}^{\tau}$, $Y = Y_* \mathrm{e}^{\eta}$, and $L = L_* \mathrm{e}^{\nu}$, where w, Y, and L are the observed quantities and τ, η, ν are random measurement errors with $E(\tau) = E(\eta) = E(\nu) = 0$.

7.4.2. *Relation Between Observations*

Maintained Hypothesis 7. Errors are statistically independent in different firms.

[Note: an alternative specification might be: (1) that ϵ follows some geographical pattern and therefore is not distributed independently over firms, (2) that the non-systematic error λ is correlated between firms, or (3) that the error in forecasting output price ξ is correlated between firms because of common output demand fluctuations.]

Maintained Hypothesis 8. Errors are homoscedastic; i.e., $\epsilon, \lambda, \xi, \eta, \nu$ have variances which do not vary across firms.

It will be necessary to make several further technical specifications in order to reach conclusions on the properties of estimators. These will be introduced as they are needed.

Taking equations (19), (22), (23), (25), plus the definitions $\tilde{w} = we^{\xi-\tau}$, $Y = Y_*e^{\eta}$, and $L = L_*e^{\nu}$, we can summarize the relations holding among the observed variables,

$$Y = AL^{\beta}e^{\epsilon+\lambda+\eta-\beta\nu}, \tag{26}$$

$$L = (we^{\xi-\tau-\epsilon}/\beta A)^{-1/(1-\beta)}e^{\nu}, \tag{27}$$

$$Y = (Ae^{\epsilon})^{1/(1-\beta)}(we^{\xi-\tau})^{-\beta/(1-\beta)}e^{\lambda+\eta}\beta^{\beta/(1-\beta)}, \tag{28}$$

$$\Pi \equiv Y - wL = (Ae^{\epsilon})^{1/(1-\beta)}(we^{\xi-\tau}/\beta)^{-\beta/(1-\beta)}\{e^{\lambda+\eta} - \beta e^{\nu-\xi+\tau}\}. \tag{29}$$

Taking logs, (26)–(29) become

$$y = \alpha + \beta l + \overbrace{\{\epsilon + \lambda + \eta - \beta\nu\}}^{u_1} = \alpha + \beta l + u_1, \tag{30}$$

$$l = \delta - \frac{1}{1-\beta}\,\omega + \overbrace{\left\{\frac{\epsilon + \tau - \xi}{1-\beta} + \nu\right\}}^{u_2} = \delta - \frac{1}{1-\beta}\,\omega + u_2, \tag{31}$$

$$y = \gamma - \frac{\beta}{1-\beta}\,\omega + \overbrace{\left\{\frac{\epsilon + \beta\tau - \beta\xi}{1-\beta} + \lambda + \eta\right\}}^{u_3} = \gamma - \frac{\beta}{1-\beta}\,\omega + u_3 \tag{32}$$

$$\begin{aligned}\pi &\doteq \theta - \frac{\beta}{1-\beta}\,\omega + \overbrace{\left\{\frac{\epsilon + \lambda + \eta - \beta\nu}{1-\beta} - \frac{\beta}{(1-\beta)^2}\frac{\tau^2}{2} - \frac{\beta(1+\beta)}{(1-\beta)^3}\frac{\tau^3}{6}\right\}}^{u_4}\\ &= \theta - \frac{\beta}{1-\beta}\,\omega + u_4,\end{aligned} \tag{33}$$

where $y = \log Y$, $l = \log L$, $\pi = \log \Pi$, $\omega = \log w$, $\alpha = \log A$, $\delta = (1/(1-\beta))\log(\beta A)$, $\gamma = (1/(1-\beta))\log A + (\beta/(1-\beta))\log\beta$, and $\theta = \gamma + \log(1-\beta)$, and where we have approximated the non-linear error in (33) by a Taylor's expansion,

$$\begin{aligned}\log\left\{\frac{e^{\lambda+\eta} - \beta e^{\nu-\xi+\tau}}{1-\beta}\right\} &\cong \frac{1}{1-\beta}\{\lambda + \eta - \beta\nu + \beta\xi - \beta\tau\}\\ &\quad - \frac{\beta}{(1-\beta)^2}\frac{\tau^2}{2} - \frac{\beta(1+\beta)}{(1-\beta)^3}\frac{\tau^3}{6} + \mathcal{O}(\lambda^2,\eta^2,\nu^2,\xi^2,\tau^4).\end{aligned}$$

The system of equations then contains the production function (30) and the first-order conditions (31). Since we deal with one input only equation (31) is also a reduced-form equation and as such it is referred to as the labor demand equation. The second reduced-form equation (32), is the supply function. Equation (33) is an approximation of the profit function.

The direct estimation of the production elasticities from the first-order conditions, termed by Klein (1953) the factor share estimate, is derived from the following relation:

$$\log \frac{wL}{Y} = \omega + l - y = \log \beta + \overbrace{(\tau - \xi + \nu - \lambda - \eta)}^{u_5}. \tag{34}$$

The first-order conditions are widely used in estimating the parameters of more complex production functions. In this case the right-hand side of

TABLE 3

No.	Name	Expression	Error structure			
			ϵ	$(\lambda + \eta)$	ν	$(\tau - \xi)$
(30)	Production	$y = \alpha + \beta l + u_1$	1	1	$-\beta$	0
(31)	First-order or labor demand	$l = \delta - c\omega + u_2$	c	0	1	c
(32)	Supply	$y = \gamma - c\beta\omega + u_3$	c	1	0	$c\beta$
(33)	Profit[a]-approximation	$\pi \doteq \theta - c\beta\omega + u_4$	c	c	$-c\beta$	0
(34)	Factor share	$\omega + l - y = \log \beta + u_5$	0	-1	1	1
(35)	First-order transformed	$\omega = (1/c)(\delta - l) + u_6$	1	0	$1/c$	1

where

$c = 1/(1 - \beta)$
ϵ = firm effect in the production function
λ = non-systematic error in the production function
η = measurement error of output
ν = measurement error of input
τ = measurement error of real wages
ξ = forecasting error of real wages

Since λ and η have the same coefficients in the various equations, the two are combined here; τ and $-\xi$ are similarly combined. The question of identifying these components is of no major concern to us and will therefore be disregarded.

[a] Add $c_2\tau^2 + c_3\tau^3$ to error term – see discussion above for details.

the equation consists of either quantities or prices. It is therefore of interest to compare these two alternatives in the present simple formulation. Such a comparison is also useful when wage is measured with an error. If such an error is more serious than the error of measuring inputs, then it may be desirable to estimate β not from (31), but rather from

$$\omega = (1-\beta)\delta - (1-\beta)l + \overbrace{\{\epsilon + \tau - \xi + (1-\beta)v\}}^{u_6}. \tag{35}$$

We refer to this as the transformed or inverted first-order condition.

In what follows it might be convenient to refer to the summary table, Table 3. The panel on the right-hand side titled "Error structure" should be read as follows: $u_1 = \epsilon + (\lambda + \eta) - \beta\nu + 0(\tau - \xi)$, and similarly for the other terms.

In order to evaluate the estimators we have to further specify the moments of the random errors. It is reasonable to assume that most of the error components are independent. The analysis begins by allowing for some non-zero covariances, as described in:

Maintained Hypothesis 9. Let $(\epsilon,\lambda,\eta,\nu,\tau,\xi) = (\cdot)$, then

$$E(\cdot) = \mathbf{0}$$

$$V(\cdot) = \begin{array}{cccccc} \sigma_\epsilon^2 & \sigma_{\epsilon\lambda} & 0 & 0 & 0 & \sigma_{\epsilon\xi} \\ & \sigma_\lambda^2 & 0 & 0 & 0 & 0 \\ & & \sigma_\eta^2 & 0 & 0 & 0 \\ & & & \sigma_\nu^2 & \sigma_{\nu\tau} & 0 \\ & & & & \sigma_\tau^2 & 0 \\ & & & & & \sigma_\xi^2 \end{array}$$

For some parts of the analysis it is also required that the first five moments of $(\cdot)$ exist.

Finally it is assumed that $0 \le \operatorname{plim}(\omega - \bar{\omega})^2 < \infty$.

The problems involved in estimation are related to the fact that the right-hand-side variables in (30)–(32) are stochastic and correlated with residuals, implying that estimates obtained by least squares (LS) will be inconsistent. We look at this more closely, and ask under what stochastic structures each estimate will be consistent. Define the sample moment about means

$$s_{yl} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(l_i - \bar{l}),$$

with corresponding notation for other moments. Table 4 presents the various estimators to be considered and their errors.

TABLE 4
Alternative estimators of β.

Estimator	Source	Error: $b_{3j} - \beta$
$b_{30} = \frac{s_{yl}}{s_{ll}}$	PF direct – by LS	$\frac{s_{l1}}{s_{ll}}$
$b_{31} = \frac{s_{\omega\omega} + s_{l\omega}}{s_{l\omega}}$	RF – ILS from labor demand	$\frac{-(1-\beta)^2 s_{\omega 2}}{s_{\omega\omega} - (1-\beta)s_{\omega 2}}$
$b_{32} = \frac{s_{y\omega}}{s_{y\omega} - s_{\omega\omega}}$	RF – ILS from product supply	$\frac{-(1-\beta)^2 s_{\omega 3}}{s_{\omega\omega} - (1-\beta)s_{\omega 3}}$
$b_{33} = \frac{s_{\pi\omega}}{s_{\pi\omega} - s_{\omega\omega}}$	Dual – ILS from approximated profit function	$\frac{-(1-\beta)^2 s_{\omega 4}}{s_{\omega\omega} - (1-\beta)s_{\omega 4}}$
$b_{34} = \exp(\bar{\omega} + \bar{l} - \bar{y})$	Factor share	$\exp(\bar{u}_5)$
$b_{35} = \frac{s_{\omega l} + s_{ll}}{s_{ll}}$	FOC – ω as "dependent variable"	$\frac{s_{l6}}{s_{ll}}$
$b_{36} = \frac{s_{y\omega}}{s_{l\omega}}$	PF direct – ω as instrumental variable	$\frac{(1-\beta)s_{\omega 1}}{(1-\beta)s_{\omega 2} - s_{\omega\omega}}$

LS = Least Squares
ILS = Indirect LS
PF = Production Function
RF = Reduced Form
FOC = First-Order Condition
$s_{\omega j} = \text{cov}(\omega u_j)$
$\bar{y}$ = sample arithmetic average of y, and similarly for other variables

Under Hypothesis 9, the probability limits of the errors of the various estimators are

$$\text{plim}(b_{30} - \beta) = \left(\frac{\sigma_\epsilon^2 + \sigma_{\epsilon\lambda} - \sigma_{\epsilon\xi}}{1-\beta} - \beta\sigma_\nu^2\right)\Big/\sigma_l^2,$$

$$\text{plim}(b_{31} - \beta) = \left(\frac{\sigma_\tau^2}{1-\beta} + \sigma_{\nu\tau}\right)\Big/\sigma_\omega^2,$$

$$\text{plim}(b_{32} - \beta) = \frac{\beta\sigma_\tau^2}{1-\beta}\Big/\sigma_\omega^2,$$

$$\text{plim}(b_{33} - \beta) = \left(\frac{\beta^2}{2(1-\beta)^3} E\tau^3 + \frac{\beta^2(1+\beta)}{6(1-\beta)^4} E\tau^4 + \cdots\right)\Big/\sigma_\omega^2,$$

$$\text{plim}(b_{34}/\beta) = 1,$$

$$\operatorname{plim}(b_{35}-\beta)=\frac{\sigma_\epsilon^2-2\sigma_{\epsilon\xi}+\sigma_\xi^2}{1-\beta}+\sigma_{\nu\tau}+(1-\beta)\sigma_\omega^2,$$

$$\operatorname{plim}(b_{36}-\beta)=(1-\beta)\beta\sigma_{\nu\tau}/\sigma_\omega^2.$$

Using these expressions we can now review the relative merits of the various estimators.

(1) Direct estimation of the production function by LS: b_{30} is a consistent estimator if the labor input is measured without error ($\sigma_\nu^2 = 0$) and there are no firm effects ($\sigma_\epsilon^2 = 0$). The specific or firm effect was first introduced in the classical paper by Marschak and Andrews (1944). As we have seen, this effect is taken into account in maximization, and consequently the input cannot be considered as exogeneous. Solutions to this problem in the framework of direct estimation of the PF were discussed by Hoch (1958, 1962), Mundlak (1961, 1963), and Mundlak and Hoch (1965). In essence, there are two basic solutions. These can be stated in a more general form that will also apply to a multiple-input technology: (i) Impose enough restrictions on the covariances of the various error terms so as to identify the production elasticities. The resultant estimate can be considered as an instrumental variable estimator where $y-l$ is the instrument; $y-l$ varies with ω and the errors which appear in line (34) of Table 3. That is, $y-l$ is obtained as a difference of the reduced-form equations, (31) minus (32). However, if λ, ν, and η are present, then the estimate obtained with $y-l$ as an instrumental variable will not be consistent. In the terminology of Mundlak and Hoch (1965), this estimator overcomes the transmitted error (ϵ) but it is susceptible to the non-transmitted errors (λ, η, and ν in our case). At this point, we digress briefly to the case of more than one input, and consider a suggestion made by Cavallo (1976). For each of the inputs there will be a first-order equation of the form of (31). Consequently, a difference between two inputs provides an instrumental variable whose systematic part consists of the price difference of the two variables and whose error part consists of the difference in measurement and forecasting errors of the two inputs, a term which does not appear in the error of the production function. Thus, the performance of such instrumental variables is independent of the relative strength of the transmitted versus non-transmitted errors. If there are k inputs, there are $k-1$ such instrumental variables and there will be a need for one more such variable. It should be noted that if there are serious errors in the measurement of inputs, such instrumental variables

will not yield a consistent estimator. (ii) The foregoing discussion is pertinent primarily to a strictly cross-section analysis. When repeated observations in time are available for each firm, covariance analysis can be applied to eliminate the firm effect, so that for the within firm variations, $\sigma_\epsilon^2 = 0$. Such an estimator is susceptible only to error in measurement of the input. Further, if the measurement error is also subject to a firm effect, then covariance analysis will solve this problem as well.

(2) Reduced-form estimates: The reduced-form estimates, b_{31} and b_{32}, require some wage rate variations among firms. Such estimates are consistent if the wage rates are measured without error. If this condition is not met, the degree of inconsistency depends on β and on the ratio $\sigma_\tau^2/\sigma_\omega^2$.

(3) Profit function: Since this equation approximates an identity in the reduced-form equations, the estimator, b_{33}, requires similar but somewhat weaker conditions for consistency than the reduced form estimators, namely that moments of third- and higher-orders be zero.

(4) Direct estimation of the production function with wages as instrumental variables: Under the present assumptions this estimator is consistent, provided measurement errors in wages and inputs are independent.

(5) Factor share: b_{34} provides a consistent estimator of β, and if the errors are log normally distributed, it is possible also to adjust the estimator so as to obtain a minimum variance unbiased estimate [Bradu and Mundlak (1970)].

(6) First-order conditions with price as a dependent variable: The consistency of this estimator requires that the input be measured without error, $\sigma_\nu^2 = 0$, that there are no firm effects in the production function ($\sigma_\epsilon^2 = 0$) and no error in optimization and price forecasting ($\sigma_\xi^2 = 0$). These are strong assumptions indeed.

Under the assumption of profit maximization the factor share estimator seems to be the simplest and easiest to compute and at the same time its consistency depends on fewer assumptions than some of the alternative estimators. This result is a direct consequence of the simplifying assumptions that the error components u_5 have zero expectations. At the purely technical level, the importance of this assumption stems from the fact that there is no other coefficient beside β to absorb deviations from such an assumption.

In the discussion of the specification of the model we have cast doubt on the general validity of the assumption $E(\xi) = 0$. Indeed, the early studies of Cobb and Douglas were largely motivated by the desire to test the hypothesis that factors are paid according to their marginal productivity. It is therefore inadequate to impose such a hypothesis as a constraint as is done in the factor share estimator [Mundlak (1963)]. This point holds equally well for more complex functional forms, the coefficients of which are estimated from the first-order conditions.

The relaxation of the assumption $E(\xi) = 0$ also has an effect on the reduced-form estimators. In evaluating the probability limit of these estimators there will be another term, $\text{plim}(\sum \omega^* \xi / n)$, and this term need not vanish even if the two components ω^* and ξ are independent. The only estimators that are not affected by this term are the direct estimates of the production function as discussed under (1) above or by using prices as instrumental variables.

We can now summarize the discussion by listing the consequences of the various error components on the alternative estimators:

(1) Firm effect in the production function (ϵ) results in inconsistency of the direct LS fit of the production function and of the transformed first-order condition estimator, b_{35}.

(2) Non-systematic error in the production function (λ) does not lead to inconsistency.

(3) Measurement errors, as is well-known, lead to inconsistency only if they occur in the independent variables of the regression. Measurement errors of the real wage, $\sigma_\tau^2 \neq 0$, lead to inconsistency of the reduced-form estimates, b_{31}, b_{32}, and possibly b_{33}. If this error is serious, it can be avoided by estimating the transformed first-order condition, b_{35}. The latter is sensitive to measurement error in input, $\sigma_\nu^2 \neq 0$, as is also the case with direct LS fit of the production function.

(4) Non-systematic errors of optimization effect only the transformed first-order condition equation.

(5) Systematic errors of optimization, which also include errors in wage forecasting, result in inconsistency of the reduced-form equations as well as the factor share estimate.

The estimator which seems to be most robust with respect to avoiding bias due to the various stochastic components is the direct estimate of the production function with real wages as an instrumental variable.

Consider now the case where there is variation in l and ω, so all the estimators b_{30} to b_{36} are defined. Suppose labor inputs and wages are measured without error ($\sigma_\nu^2 = \sigma_\tau^2 = 0$). Which of the estimators is "best"? A partial answer for large samples can be obtained by comparing asymptotic variances. A tiresome computation for the case of normally distributed errors yields

$$\begin{aligned}\text{plim } n(b_{30} - \text{plim } b_{30})^2 = {} & (1-\beta)^2\{\sigma_\epsilon^2 + \sigma_\lambda^2 + \sigma_\eta^2 + 2\sigma_{\epsilon\eta})\sigma_\omega^2/(\sigma_\omega^2 + \sigma_\epsilon^2 \\ & + \sigma_\xi^2 - 2\sigma_{\epsilon\xi})^2\} + (1-\beta)^2\{(\sigma_\epsilon^2 - 2\sigma_{\epsilon\xi} \\ & + \sigma_\xi^2)(\sigma_\lambda^2 + \sigma_\eta^2) + (3\sigma_\epsilon^4 - 6\sigma_\epsilon^2\sigma_{\epsilon\xi} + \sigma_\xi^2\sigma_\epsilon^2 \\ & + 2\sigma_{\epsilon\xi}^2\}/(\sigma_\omega^2 + \sigma_\epsilon^2 + \sigma_\xi^2 - 2\sigma_{\epsilon\xi})^2, \qquad (37)\end{aligned}$$

$$\text{plim } n(b_{31} - \beta)^2 = (1-\beta)^2\{\sigma_\epsilon^2 + \sigma_\xi^2 - 2\sigma_{\epsilon\xi}\}/\sigma_\omega^2,$$

$$\begin{aligned}\text{plim } n(b_{32} - \beta)^2 = {} & (1-\beta)^2\{\sigma_\epsilon^2 + \beta^2\sigma_\xi^2 - 2\beta\sigma_{\epsilon\xi} \\ & + (1-\beta)^2(\sigma_\lambda^2 + \sigma_\eta^2)\}/\sigma_\omega^2,\end{aligned}$$

$$\text{plim } n(b_{33} - \beta)^2 = (1-\beta)^2\{\sigma_\epsilon^2 + \sigma_\lambda^2 + \sigma_\eta^2 + 2\sigma_{\epsilon\lambda}\}/\sigma_\omega^2,$$

$$\text{plim } n(b_{34} - \beta)^2 = \beta^2\{\sigma_\xi^2 + \sigma_\lambda^2 + \sigma_\eta^2\},$$

$$\begin{aligned}\text{plim } n(b_{35} - \text{plim } b_{35})^2 = {} & (\sigma_\epsilon^2 + \sigma_\xi^2 - 2\sigma_{\epsilon\xi})\{\sigma_\omega^2 + 2(\sigma_\epsilon^2 + \\ & \sigma_\xi^2 - 2\sigma_{\epsilon\xi})\}/(1-\beta)^2\sigma_l^2,\end{aligned}$$

$$\text{plim } n(b_{36} - \beta)^2 = \frac{(1-\beta)^2(\sigma_\epsilon^2 + \sigma_\lambda^2 + 2\sigma_{\epsilon\lambda}) + 3\beta(1-\beta)\sigma_{\epsilon\xi} + 4\beta^2\sigma_\xi^2}{\sigma_\omega^2 + \sigma_\epsilon^2 + \sigma_\xi^2 - 2\sigma_{\epsilon\xi}}.$$

When $\sigma_\epsilon^2 = 0$, so that b_{30} is consistent, (37) reduces to

$$\text{plim } n(b_{30} - \beta)^2 = (1-\beta)^2(\sigma_\lambda^2 + \sigma_\eta^2)/(\sigma_\omega^2 + \sigma_\xi^2).$$

In this case, the relative efficiency of the estimators depends on relative variances. For example, if optimization errors (σ_ξ^2) are large, then b_{31} and b_{32} are undesirable and b_{30} will tend to be most efficient. If σ_ξ^2 is low, then b_{31} will tend to be most efficient and b_{32} will be more efficient than b_{33}. When σ_ω^2 is low relative to σ_ξ^2, b_{30} will be most efficient. For β near one, the estimators b_{30} to b_{33} will be relatively efficient, while β near zero will make b_{34} most efficient. When σ_ϵ^2 is large, b_{34} will tend to be most efficient. All these conclusions, it should be noted, are shown only for large samples. While it is dangerous to over-generalize from specialized small-sample results, there seems to be a tendency for direct ordinary least-squares estimators such as b_{30} to be the best estimators in small samples more often than one would guess from the extrapolation of asymptotic results; i.e., direct LS estimators seem to be somewhat more robust than their competitors in small samples. The exact small sample distributions of the estimators b_{31} to b_{33} can be derived for the example above, and are found to have tails that behave like Cauchy distributions. Consequently, the mean and variance of b_{31} to b_{33} are not defined in finite samples, and the probability of estimates of β which are far from the true value are rather large. Thus, the estimators in this example tend to confirm the generalization regarding the relative small sample robustness of direct ordinary least-squares estimators.

Before concluding the discussion it should be pointed out that we have made repeated reference to the instrumental variable estimator. Such an estimator overcomes difficulties caused by measurement errors and lack of independence between the explanatory variables and the error terms. In general there are several difficulties with the use of this method of which the user should be aware. First, instrumental variable estimates are not as efficient (have larger variance) as the direct LS estimator. This problem can be reduced by a proper selection of instrumental variables, which leads us to the second problem – that of finding such variables. Instrumental variables should be uncorrelated with the error terms in the equation and at the same time be correlated with the explanatory variables. The larger is the latter correlation (properly defined when there are more than one variable) – the smaller is the variance of the estimator. Third, in small samples, instrumental variables estimators usually have distributions with "fat" tails, tending to produce extreme values. Thus, one may buy consistency at the cost of a less accurate estimator in a small sample.

In the foregoing discussion we considered the use of the real wage as an instrumental variable. This generalizes to the use of real factor prices in the case of more than one input and the use of product price ratios in the case of a multiproduct production function. We have also mentioned the use of some linear combinations of the quantities as instrumental variables which can eliminate some of the errors. All these are variables which come from the model. It is also possible to use variables which come from outside the model [Berndt and Christensen (1973a)].

We have discussed in the context of the example above the difficulties encountered when there is insufficient variation in the independent variables in a direct or indirect estimation of production parameters. In the case of multiple inputs, this problem reappears as that of *multicollinearity*, or high correlation among the independent variables so that there is insufficient cross-variation to allocate with precision the contribution of separate variables to the determination of the dependent variable. This problem is particularly acute for time series analysis, and in functional forms where the independent variables appear as "substitutes". Considerable success in dealing with multicollinearity has been achieved in production applications by considering the production function as part of a complete economic model. In the example above, one may view (30)–(33) as equations in a simultaneous system. When non-redundant sets of equations are estimated jointly, they can provide more efficient estimates than any one equation considered above. The effectiveness of the analysis of complete systems to reduce multicollinearity and increase precision is most evident in estimation of general linear-in-parameters forms such as the Diewert or translog systems [e.g., Burgess (1975) and Woodland (1975)].

We can summarize our conclusions including the inferences that can be drawn from the example. The relative desirability of estimation of the production function, its dual profit or cost function, factor demand or supply equations, or their inverse first-order conditions, depends primarily on the stochastic structure of the data. In the general case, these equations together constitute a simultaneous system, and the most efficient estimators are obtained by estimation of the complete system. When the source of stochastic errors is confined to technological effects not observed by production units, then direct estimation of the production function is a good procedure, although multicollinearity will be a problem for many data sets. In principle, estimation of factor demand equations in which the independent variables are prices is a good procedure, being consistent in the presence of stochastic components which make direct estimation of the production function inconsistent.

However, if multicollinearity constitutes a problem in the direct estimation, it is likely to remain so in the estimation of the factor demand equations. This problem is overcome in part by estimation of the first-order conditions, which under the separability conditions frequently imposed in empirical analysis have fewer variables than the demand functions. The use of the first-order equations represents, like the direct estimation of the production function, a limited information approach which does not use all the constraints of the system.

It should be noted that several caveats apply in the use of dual profit or cost functions and their derivative demand and supply functions, as well as first-order conditions. First, the construction of these functions require maintained hypotheses on market environment and behavior which may not be necessary for direct estimation of the production function. Failure of one of these maintained hypotheses may result in a model which does not have the postulated structural relationship to the underlying technological parameters. For example, if markets are not competitive, or if firms fail to maximize profits, non-technological factors are introduced into the "as if" technology reconstructed under a competitive profit maximization assumption. Second, there may be insufficient variation in factor prices to allow accurate estimation of production parameters. Mundlak (1968a) has noted that variation in production quantities in many data sets is much greater than variation in prices. This is presumably due to random effects on technology, environment, or firm behavior. In some cases, this may mean that more accurate estimates can be obtained by direct estimation, even in the likelihood of an introduction of bias. On the other hand, McFadden (Chapter IV.1) has found in a data set on establishments substantial price variation at the plant level in inputs which have "national" markets, due to transportation costs, timing of purchase, volume of contracts, and local conditions. This suggests that indirect methods may be quite satisfactory when accurate establishment price data are available, but may perform poorly when more general market price indices are used.

Third, in the analysis of firms facing non-competitive markets, or of industry or macroeconomic production aggregates, prices are not exogenous. Valid estimation requires information of the remainder of the system, with simultaneous estimation; or the use of instrumental variables methods. As noted in the example, the small sample advantages of ordinary least-squares regression estimates over instrumental methods may suggest use of instrumental estimators only for large data sets.

8. Overview of Empirical Analysis

The empirical literature which utilizes Cobb–Douglas and CES production functions has been surveyed by Walters (1963), Nerlove (1967), and Bridge (1971). The outstanding example of the use of the Cobb–Douglas cost function is Nerlove's (1963) study of electricity supply. The Cobb–Douglas profit function has been used by Lau and Yotopoulos (1971) to analyze efficiency in Indian agriculture. An example of the use of a CES cost function can be found in Chapter IV.1 of this volume by McFadden. Recent estimates of Cobb–Douglas and CES production functions can be found in Griliches and Ringstad (1971).

In the past several years most of the empirical literature has been devoted to attempts to implement the flexible functional forms discussed in Section 4. Generalized Leontief cost functions have been estimated for Sweden by Parks (1971), for Canada by Woodland (1975), and for

Norway by Frenger in Chapter V.2 of this volume. Fuss, Chapter IV.4, estimated a two-stage nested variant of this function for the U.S. steam-electric generation industry. Translog production functions have been applied to U.S. manufacturing by Berndt and Christensen (1973a, 1974) and to aggregate U.S. activity by Christensen, Jorgenson, and Lau (1973) and Burgess (1975). Examples of the estimation of translog cost functions are papers by Burgess (1975), Denny and Pinto (Chapter V.1), Berndt and Wood (1975), and Fuss (1977a). Translog profit functions have been utilized by Christensen, Jorgenson, and Lau (1973), and Hudson and Jorgenson (1974). Finally, the quadratic profit function is the functional form used in Cowing's study of the regulatory constraint, Chapter IV.5 of this volume.

9. Conclusion

This chapter has stressed the importance of economic and statistical criteria for the choice of functional forms in the estimation of production relationships. We have pointed out that linear-in-parameters forms provide a flexible, general purpose approach to functional specification, and that the linear-in-parameters approach can be utilized to tailor functional forms to specific applications. However, we have also emphasized the use of non-linear functional forms in applications where economy and ease of interpretation of parameters is important, as in studies of elasticities of substitution. The critical role of separability as an economic assumption, and as a tool in the construction of functional forms, has been stressed. Finally, we have used a simple example to illustrate the implications of alternative sources of stochastic error for the choice of functional form and estimation method.

We emphasize in conclusion that the primary interest in specific functional forms lies in their empirical application, and that the choice of a functional form should be based on an integrated consideration of the economic problem and likely stochastic structure of the observed data.