HANDBOOKS IN ECONOMICS 18

HANDBOOK OF AGRICULTURAL ECONOMICS

VOLUME 1A
AGRICULTURAL PRODUCTION

Editors:
Bruce L. Gardner
Gordon C. Rausser

NORTH-HOLLAND

INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments, from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

KENNETH J. ARROW and MICHAEL D. INTRILIGATOR

PUBLISHER'S NOTE

For a complete overview of the Handbooks in Economics Series, please refer to the listing on the last two pages of this volume.

CONTENTS OF THE HANDBOOK

VOLUME 1A

PART 1 – AGRICULTURAL PRODDCTION

VOLUME 1B

PART 2 – MARKETING, DISTRIBUTION AND CONSUMERS

INTRODUCTION

The subject matter of agricultural economics has both broadened and deepened in recent years, and the chapters of this Handbook present the most exciting and innovative work being done today. The field originated early in the twentieth century with a focus on farm management and commodity markets, but has since moved far into analysis of issues in food, resources, international trade, and linkages between agriculture and the rest of the economy. In the process agricultural economists have been pioneering users of developments in economic theory and econometrics. Moreover, in the process of intense focus on problems of economic science that are central to agriculture – market expectations, behavior under uncertainty, multimarket relationships for both products and factors, the economics of research and technology adoption, and public goods and property issues associated with issues like nonpoint pollution and innovations in biotechnology – agricultural economists have developed methods of empirical investigation that have been taken up in other fields.

The chapters are organized into five parts, contained in two volumes. Volume 1 contains Part 1, "Agricultural Production", and Part 2, "Marketing, Distribution and Consumers". These two parts include much of the traditional scope of agricultural economics, emphasizing advances in both theory and empirical application of recent years. Volume 2 consists of three parts: "Agriculture, Natural Resources and the Environment", "Agriculture in the Macroeconomy", and "Agricultural and Food Policy". Although agricultural economists have always paid attention to these topics, research devoted to them has increased substantially in scope as well as depth in recent years.

A large-scale effort to review and assess the state of knowledge in agricultural economics was previously undertaken by the American Agricultural Economics Association (AAEA), with publication in four volumes from 1977 to 1992.[1] Those earlier survey volumes have strikingly different subject-matter content from that of the present Handbook, especially considering that they described the same field only 20 years ago. The AAEA volumes have extensive coverage of farm management issues, costs of production in agriculture, and estimates of efficiency of marketing firms. In our judgment little in any fundamental way has been added to our knowledge in these areas, and applications have become routine rather than imaginative research. The largest AAEA volume was devoted entirely to agriculture in economic development. This remains a

[1] *A Survey of Economics Literature*, Lee Martin, ed., Minneapolis: University of Minnesota Press. Volume 1, Traditional Field of Agricultural Economics (1977); Volume 2, Quantitative Methods in Agricultural Economics (1977); Volume 3, Economics of Welfare, Rural Development, and Natural Resources (1981); Volume 4, Agriculture in Economic Development (1992).

most important topic, but we cover it in only one complete chapter and parts of several others. This reflects in part the integration of work on developing countries with mainstream applied work. For example, our chapters on production economics, expectations, and risk management also encompass applications to agriculture in developing economies.

That integration points to another gradual but notable change in agricultural economists' research. The AAEA surveys had most of the chapters of one volume devoted to quantitative methods. We do not have any separate methodological chapters. In contrast, we have several chapters with substantial development of economic theory. This reflects an evolution in the research priorities of leading agricultural economists who, following the earlier work of Nerlove on supply and Griliches on technological change, are working at the theoretical frontiers and simultaneously undertaking empirical work – not just purveying new theories to their more "applied" colleagues.

As its title indicates, the AAEA volumes were surveys of literature, and aimed at completeness of coverage within their subject matter. We asked our authors to be selective, to focus on what they saw as the main contributions to the area they covered, and to assess the state of knowledge and what remains to be learned. This approach has left some gaps in our coverage, and has given us some chapters that are perhaps more idiosyncratic than is usual for a survey chapter. In order to pull things together at a higher level of aggregation, we commissioned five "synthesis" chapters, one for each of the five parts of the Handbook. And, to provide our own even broader overview, the editors have written closing syntheses of each volume. Because these syntheses provide capsule summaries of each Handbook chapter, we will not present further description of content here.

Although advances in research in agricultural economics are increasingly being made in many countries, our authors and coverage of applied topics is heavily U.S.-weighted (only six authors work outside of the U.S.: two in Europe, two in Australia, one in Canada, and one in Israel). Of those in the U.S., however, six are economists at the World Bank, an international rather than American institution. Probably in another twenty years or so one will have to become more international to capture the most interesting and exciting developments in the field, but that day has not arrived yet.

Among the many debts we have accrued in the preparation of this Handbook, the most important was Rachael Goodhue. She not only assessed the substance of many chapters, but she persuaded many reviewers and authors alike to complete their assigned responsibilities. Other critical contributors include the dedicated staff who provided support at the University of California, Berkeley, and at the University of Maryland. At Maryland, Liesl Koch served as copy editor and guided the authors' final revisions and preparation of the manuscript with sure judgment and a firm but diplomatic hand, a job best likened to driving a herd of cats. Coordination of correspondence with authors and reviewers was organized and carried out at Berkeley with exemplary efficiency and organizational skill by Jef Samp, Jessica Berkson, and Jennifer Michael, under the direction of Nancy Lewis.

We also want to recognize the comments and suggestions received from 45 reviewers of chapter drafts: Julian Alston, Jock Anderson, Richard Barichello, Eran Beinenbaum, Michael Boehlje, Dan Bromley, Steve Buccola, Allan Buckwell, David Bullock, Michael Caputo, Jean-Paul Chavas, John Connor, Klaus Deininger, Jeffrey Dorfman, Marcel Fafchamps, Gershon Feder, Joe Glauber, Dan Gilligan, Rachael Goodhue, Tom Grennes, Zvi Griliches, Geoff Heal, Eithan Hochman, Matt Holt, Wallace Huffman, D. Gale Johnson, Zvi Lerman, Erik Lichtenberg, Ethan Ligon, Alan Love, Jill McCluskey, Mario Miranda, Arie Oskam, Dick Perrin, Mark Rosegrant, Vern Ruttan, Ed Schuh, Kathleen Segerson, Larry Sjaastad, Spiro Stefanou, Jo Swinnen, Frans van der Zee, Finis Welch, Abner Womack, and Jacob Yaron.

BRUCE GARDNER
GORDON RAUSSER

Chapter 1

PRODUCTION AND SUPPLY

YAIR MUNDLAK

Faculty of Agriculture, The Hebrew University of Jerusalem, Rehovot, Israel

Contents

Handbook of Agricultural Economics, Volume 1, Edited by B. Gardner and G. Rausser

Abstract

The work of more than 50 years aimed at gaining empirical insight into the production structure of agriculture and the related modes of farmers' behavior is reviewed, and orders of magnitude of the various parameters of interest are quoted. The review follows the lines of the evolution of the pertinent research, and it builds on it in forming a general framework for empirical work. This approach broadens the scope of producers' decisions to include the choice of the implemented technology and it also overcomes statistical problems that have accompanied the relevant research for a long time.

JEL classification: Q11

Technology along with the competitive conditions constitute the core of the supply side of the economy. There is hardly a subject in economics that can be discussed with production sitting in the balcony rather than playing center stage. To mention the main favorable subjects in agricultural economics research: product supply, factor demand, technical change, income distribution, the relationships between factor prices and product prices, the competitive position of agriculture, returns to scale, the size distribution of firms, and capital accumulation. The nature of the relationships and the conclusions derived in any particular analysis depend on the order of magnitude of the parameters in question. Hence, whether we want it or not, the empirical analysis of technology and its changes is of cardinal importance, and measurement problems are pertinent even if on the surface it seems that the subject matter is not 'technical'.

In this review, we deal with the various aspects of the analysis. As will become clear, much of the discussion in the literature is methodology driven, not always accompanied by substantive applications. Inasmuch as methodological innovations are desirable, the question is how do they help us to think of, or deal with, specific issues of interest. This is a question that the reader should try to answer for himself, depending on his particular interest. To assist in this endeavor, we summarize here the empirical findings that bear on the main parameters of interest and address some important methodological issues essential to the interpretation of empirical studies and to future research. In many cases, the empirical results display a wide range and thus highlight the need for an appropriate framework for their evaluation. The choice of subjects and the coverage in the discussion are carried out with the purpose of constructing a uniform framework to meet the purpose. This is built on the cumulative experience and contributions provided by numerous studies and on the evolution of the thinking that is so valuable in the reading and the interpretation of the data. To emphasize this aspect, the subjects are introduced largely in an order that highlights this evolution.

There are two fairly distinct periods in the study of agricultural production functions: before and after duality. The changing of the guard was in the early 1970s, although a few studies employing direct estimation continue to appear after 1970. The appearance of duality changed not only the method of estimation but also the questions asked to the extent that there is little continuity in the subjects of interest. This can be accounted for by the fact that much of the work is methodology-driven rather than being an indication that the old questions had been adequately answered or of any explicit agenda.

1. Primal estimates or the Cobb–Douglas culture

1.1. The setting of the agenda

It seems that the empirical work on agricultural production functions originated in a methodological paper by Tintner (1944) and an application by Tintner and Brownlee (1944), which appeared as a short paper in the Notes section of the *Journal of Farm*

Economics and was followed by a full size paper by Heady (1946). This work was influenced by the work of Cobb and Douglas (1928).[1] It thus took about fifteen years to adopt the work of Cobb and Douglas in agricultural economics application.

These studies used data from a random sample of Iowa farms for 1939. The data were classified by area of the state, type, and also size of farm. The inputs included were land, labor, equipment, livestock and feed, and miscellaneous operating expense, a classification that is still applicable today. Interestingly, this early work anticipated some of the more difficult subjects in the empirical work of production functions. Management was recognized as an input, but "[t]he productive agent management has been excluded since there is no satisfactory index of inputs for this factor" [Tintner and Brownlee (1944, p. 566)]. Allusions were also made to the importance of input quality.[2] Heady (1946) expressed similar concerns about the quality issue and the omission of management.[3] Also, based on the criticism of the Cobb–Douglas work that appeared at that time by Reder (1943), Bronfenbrenner (1944), and Marschak and Andrews (1944), Heady (1946) noted that "[t]he functions which have been derived ... are of the inter-firm rather than intrafarm variety ... it can be expected that a multitude of functions exists ... because of the varying combinations of techniques employed and commodities produced" (p. 999). This is a recognition of the problems caused by aggregation over techniques. Similarly, Smith (1945) observed that firms in cross section may employ different techniques, particularly due to fixed plants inherited from the past, and the long-run production functions so derived may represent "mongrels" or hybrids. Aside from the question of input quality, Bronfenbrenner (1944) raised the point that capital and labor are not on the same footing because labor is a flow ("quantity used"), whereas capital is a stock (representing the "available quantity"). This can be interpreted as an early recognition of the conceptual problem of the evaluation of the productivity of durable inputs.

These studies were concerned with the contribution of inputs to output variations and with a comparison of the factor productivity on different farm types and the relationship to their returns. The estimated production elasticities reported by Tintner and Brownlee (1944) for the sample as a whole are: land, 0.34; labor, 0.24; and other assets and variable inputs, 0.41. The sum is 0.99. Heady used a larger sample and a somewhat different classification of inputs to obtain for the sample as a whole: land, 0.23; labor, 0.03; and other assets and variable inputs, 0.59. The sum is 0.85.

[1] A regression equation linear in the logarithms "[is] similar to the production function employed by Paul Douglas in his empirical studies" [Tintner and Brownlee (1944, p. 567)]. On the history of the Cobb–Douglas production function, see [Douglas (1976)].

[2] "Using the number of acres in the farms as a measure of inputs of land ignores variations in the quality of land. Measuring inputs of labor in terms of months of labor also ignores variations in the quality and intensity of labor, particularly that of operator and his family" [Tintner and Brownlee (1944, p. 566)].

[3] At the time the issue of management bias was unrecognized, therefore both papers speculated that had management been included, the sum of the elasticities, as a measure of returns to scale, would have increased [Tintner and Brownlee (1944, p. 569), Heady (1946, p. 995)]. However, Heady also indicates that the sum of the elasticities might have decreased due to the introduction of management (Ibid., p. 997).

Several points are of interest. First, these studies were prompted by a methodological innovation introduced by Cobb and Douglas (1928). Yet, their orientation is applicative in nature, and they address substantive issues related to the efficient use of inputs. Second, sampling from the same data source yields different elasticities. The sum of the elasticities of labor and land vary between 0.58 and 0.25 in the two studies respectively. This difference suggests sensitivity of the estimates to output composition and perhaps differences in the physical environment. Third, the sum of the elasticities is smaller than 1.

The approach formulated by the foregoing studies served as a framework for the production function estimation for more than two decades, where attention was focused on the following issues: the contribution of the various factors to the explanation of output variations in the cross section or over time, the production elasticities and their significance, the robustness of the estimates, the role of economies of scale, as judged by the sum of the elasticities, the importance of the quality of inputs, the treatment of management and its relations to the properties of the estimates, the functional forms, and the role of technical change. The data base of these studies varied from observations on individual farms to cross-country comparisons.

The question of efficient use of inputs is the objective of many studies.[4] Lack of robustness of empirical results was raised by Hildebrand (1960) who found that annual cross-section regressions are not robust and any hypothesis can be supported by some results. Lack of robustness is also evident in some other studies that present more than one set of results. Heady and Dillon (1961, Chapter 17) review and summarize 32 studies in various countries based on farm data. The mean elasticities and their coefficient of variation (in parentheses) are: land 0.38 (0.58), labor 0.21 (0.80), and "other services" 0.39 (0.59). In all these studies the sum of all the elasticities is near 1. The magnitude of the coefficient of variation indicates a wide spread in the results among the studies. They compare their results with those obtained in the pioneering cross-country study by Bhattacharjee (1955) and with assumptions made in the literature.[5] All of this indicates an effort to get a definitive substantive solution. But as this target was realized to be elusive, they concluded that "[s]till, the variations shown among the elasticities of Table 17.14 bear witness to the dangers associated with the use of any such global production function" [Heady and Dillon (1961, p. 633)].[6] The discussion is then shifted to the examination of the efficiency of the resource use. For instance, their Table 17.17 presents a ratio of the marginal productivity of labor to its opportunity cost with values varying between 2.84 observed in Taiwan to negative values obtained in dairy farming in Sweden. The median value of this ratio is 0.67. They present similar calculations for land

[4] See, for instance, Hopper (1965), Chennareddy (1967), Sahota (1968), and Herdt (1971) for India; Yotopoulos (1967) for Greece; Huang (1971) for Malaya; and Headley (1968) for the US.

[5] Bhattacharjee (1955, regression 4) reports elasticities of 0.36 and 0.3 for land and labor respectively.

[6] Clark (1973) assembles many results of factor shares in an informal framework but with good international coverage. It is very clear that the estimates depend on the economic environment which is a major theme of our discussion.

and capital services, but these are more problematic for conceptual reasons which need not be discussed at this point. To get a view of the diversity of the results, the reader is advised to check some of the country studies based on the primal approach.[7]

In 1944 Marschak and Andrews pointed out that the inputs are endogenous, and therefore Ordinary Least Squares (OLS) estimates of the production function are biased. Their paper extended the scope of the analysis by introducing issues related to the statistical properties of the estimates. Their work and Haavelmo's (1947) work on the consumption function were early examples of the problems of simultaneity in economic analysis and thus revived the question that had been asked by Working (1927) about the meaning of statistical demand equations. That opened up a route of work centered on methodological issues with a life of its own.[8]

The simultaneity problem in the estimation of production functions was overcome by the factor share estimator proposed by Klein (1953) and applied by Wolfson (1958). This estimator is based on the assumption that firms always employ *all* their inputs so as to satisfy the first order conditions for profit maximization given the *current ex post* prices. As such, the factor share estimator is subject to a major conceptual difficulty in that it cannot answer the original question of Cobb and Douglas about the empirical relevance of the competitive conditions because they are imposed in the derivation of the estimator.[9] Although this is seldom explicitly recognized, or acknowledged, all the estimators that use the first order conditions for profit maximization – and to be sure, these include the estimators based on duality as well as on the axioms of revealed preferences – use the very same property and thus are subject to the same limitation.

A different line of attack on the simultaneity problem was taken by Mundlak (1961) and Hoch (1962) through the use of covariance analysis.[10] Applying this method to a sample of family farms in Israel gave lower estimates for the elasticities compared

[7] For instance, in addition to the studies mentioned in footnote 5, US: Tintner and Brownlee (1944), Heady (1946), Hildebrand (1960), Griliches (1963a, 1963b, 1964), Kislev (1966), Tweeten and Quance (1969), Kislev and Peterson (1996); India: Lau and Yotopoulos (1972); Israel: Mundlak (1961), Sadan (1968); Mexico: Ulveling and Fletcher (1970); Colombia: Colyer and Jimenez (1971); Taiwan: Yotopoulos, Lau, and Lin (1976), Shih, Hushak, and Rask (1977), Wu (1977); Thailand: Mittelhammer, Young, Tasanasanta, and Donnelly (1980).

[8] The early work on production functions, up to the early 1960s, is surveyed by Walters (1963).

[9] I found the following statement by Clark (1973, fn 8, p. 21) to be interesting: "Douglas told me that when the function was first prepared in the 1920s, he was expecting it to show that wages then actually received by labour were considerably below its true marginal product; and was surprised to find that they were in fact extremely close to the level predicted by the function".

[10] Hoch (1958) examined a solution to the simultaneity problem based on identification through the second moments of the equations disturbances. There is no reference in the literature to an empirical application of this method, perhaps for a good reason because, as indicated by Mundlak and Hoch (1965), it is very sensitive to the specification and in the case of a likely specification error can have an unbounded bias. In another paper, Hoch (1955) suggested the use of covariance analysis. However, the method was not discussed in connection with the simultaneity problem. This is probably the reason that covariance analysis was not mentioned in [Hoch (1958)], which deals head-on with that problem. It is only in [Hoch (1962)] that the covariance analysis is seen as a solution to the simultaneity problem.

to OLS without allowance for firm effect, and their sum declined from roughly 1 to roughly 0.8. Mundlak (1961) interpreted the difference between 1 and the sum of the elasticities as the factor share of management.[11] The method was also used to estimate the managerial capacity and its empirical distribution in [Mundlak (1964a)]. Another substantive result of that study is an elasticity of land near zero. The farms in the sample are very small, and on the surface one would have expected a higher elasticity for land. However, a low elasticity for land is indicative of low profitability of agriculture. This interpretation is supported by the fact that a negligible elasticity for land in Israel was also obtained for a sample of large farms (kibbutzim) in [Sadan (1968)], so the result is unrelated to farm size.

The observations made so far are:

O.1 The estimates are not robust.
O.2 Often, results show a gap between marginal productivity and real factor prices.
O.3 Specifically, there is a difference between estimates based on inter and intrafarm observations.
O.4 Firms use different techniques.
O.5 Input quality is not addressed.
O.6 A lack of clarity on whether to use stock or flow variables.
O.7 Inputs are endogenous, and therefore OLS estimates are inconsistent.
O.8 It is possible to overcome the problem of inconsistency.
O.9 A need to further explore the role and scope of factor-share estimates.

1.2. A simple production model

The initial discussion can be conducted in terms of a single-input Cobb–Douglas production function

$$Y = AX^{\beta}\,\mathrm{e}^{m_0+u_0}, \tag{1}$$

where m_0 is the firm effect, or management, a firm-specific factor known to the firm but not to the econometrician (private information), and u_0 is a random term whose value is not known at the time the production decisions are made. The conditional expectation of output, given the input, of firm i is[12]

$$Y_i^e \equiv \mathrm{E}(Y|_{X_i}) \cong AX_i^{\beta}\,\mathrm{e}^{m_{0i}}. \tag{2}$$

[11] Other sources of farm-specific effects are differences in land quality, micro-climate, and so on. However, the emphasis has been placed on management. The firm effect is observed not only in production functions estimated from farm data; it is also a common phenomenon in cross-section analysis of manufacturing data. Thus, it seems that differences due to farming environment are not the main reason for the firm effects.

[12] Note that $\mathrm{E}(\mathrm{e}^{u_0}) \cong (1+\sigma_{00}^2/2)$; $\sigma_{00}^2 = \mathrm{E}(u_0^2)$. This term is ignored in (2).

At this stage we assume that the price is known, and the firm chooses the input so as to maximize the expected profit:

$$\max_{X_i} \pi^e(X|_{W,P,i}) = PY_i^e - WX_i, \tag{3}$$

where P and W are the product and input prices respectively. The first order condition is met up to the stochastic terms m_1 and u_1

$$\beta A X^{\beta-1} = \frac{W}{P} e^{m_1+u_1}, \tag{4}$$

where m_1 is known to the firm but not to the econometrician, and u_1 is a transitory component. The term m_1 reflects the firm's expectation formation and its utility function. In what follows, we will deal with real prices, so that W is the wage in output units, and P is the product price in input units.

We write Equations (2) and (4) in logarithms, with the variables measured as deviations from their overall mean, and introduce time notations:

$$y_{it} - x_{it}\beta = m_{0i} + u_{0it}, \tag{5}$$

$$y_{it} - x_{it} = w_{it} + m_{1i} + u_{1it} + u_{0it}. \tag{6}$$

When prices are exogenous the reduced form for x (note that $p = -w$) is

$$x_{it} = -c(p_{it} + u_{1it} + m_{1i} - m_{0i}); \quad c = (1-\beta)^{-1}. \tag{7}$$

The four error components are assumed to be IID with the following first two moments:

$$u_{jit} \sim (0, \sigma_{jj}); \quad m_{ji} \sim (\mu_j, \tau_{jj}); \quad j = 0, 1, \tag{8}$$

where $\mu_0 = 0$ and μ_1 is unrestricted. The expected value of all cross products of the error components is zero.[13]

Several of the observations made above are related to the endogeneity of the input. Equation (7) shows that the input is a function of the firm effect, m_{0i}, which is also part of the production function shock, and therefore the input is not exogenous. The bias caused by this dependence contributes to the lack of robustness. Specifically, it contributes to the differences between intra and interfirm estimates (O.3). Also, when biased coefficients are used to test the efficiency of resource use, an erroneous conclusion of an inefficient use of resources (O.2) might be reached even when the firms use resources efficiently, or conversely.

[13] Shocks that affect all firms generate time effects that can be treated in the same way as the firm effect. The extension to include time effects is straightforward and need not be reviewed here (see [Mundlak (1963a)]).

Several approaches are offered to overcome the problem of input endogeneity (O.7). When the sample consists of panel data, covariance analysis transforms the variables to deviations from the firm mean, and thereby the firm effect is eliminated from Equation (7). Let the sample average over the time observations be x_i; then Equation (7) is transformed to

$$x_{it} - x_{i.} = -c(p_{it} - p_{i.} + u_{1it} - u_{1i.}), \tag{9}$$

and it is seen that the firm effect has disappeared. The estimator is referred to as a "within" estimator (because it is based on within-firm variations).

An alternative approach is to use the price as an instrumental variable for estimating Equation (5). This is basically the dual approach to estimation, to be discussed below. This estimator is likely to be less efficient than the covariance estimator because it does not use all the pertinent information [Mundlak (1996a)]. This can be seen intuitively from Equation (7). The variability of the input in the sample is generated by four components: p_{it}, u_{1it}, m_{1i}, and m_{0i}. The last term causes the bias and should be eliminated, whereas the other three terms provide the information for the estimation. Hence, the most efficient procedure would be to use the first three components as instrumental variables. However, this cannot be done directly because, of the three variables, only p is observed. The within estimator uses the within-firm variations of p and u_1 as instruments, whereas the dual estimator uses as an instrumental variable the total variations of p but does not utilize the information in u_1. The point is that any variability of input, regardless of whether or not it is consistent with the first order condition for profit maximization, generates points on the production function and therefore helps to trace it, or more technically, helps to identify the production function.

The use of price as an instrument is subject to some limitations. If the sample consists of competitive firms, the between variability of the prices should be nil. If the sample consists of market (rather than micro) data, then the prices are not necessarily exogenous and therefore cannot be used as instrumental variables. In any case, it is possible to combine the two estimators by using the within-input variable and the price as two instrumental variables. Other possible modifications are suggested in [Mundlak (1996a)]. However, all these have not been tried out. The empirical experience is limited to the 'within' and the dual estimators. Some of the results with respect to the 'within' estimator have been mentioned above, whereas the empirical experience with the dual estimator will be discussed below.

The factor-share estimator imposes the first order conditions for profit maximization, in which case the factor share is equal to the production elasticity, β, up to a stochastic term. Using Equation (6) it is easy to see that this estimator is inconsistent.

An important issue in the empirical investigation is whether the function displays constant returns to scale (CRT). If it does, in the case of the single-input function, β is equal to 1, and there is nothing to estimate. Thus the problem is more pertinent to the more realistic case with more than one input. To see this, assume now that there are k inputs. In this case, the model consists of Equation (5) where x and β will be

k-vectors and k-equations of the form of (6) [Cavallo (1976)]. Note that the difference of the first-order conditions for any two inputs, say 1 and 2, is free of m_0 and of u_0

$$x_2 - x_1 = w_2 - w_1 + u_2 - u_1 + m_2 - m_1. \tag{10}$$

Therefore, $x_2 - x_1$ can serve as an instrumental variable. Note that this variable contains all the pertinent information related to the two inputs. There are $k - 1$ such instruments, and there is a need for one more instrument to complete the estimation of the system. The assumption of CRT is a good candidate. In this case, a Cobb–Douglas function where the variables are divided by one of the inputs is free of simultaneous-equations bias.

1.3. Productivity

To understand some of the subsequent literature we turn to another direction of inquiry, that of measuring factor productivity, that was taking place at the same time. The most influential work in agriculture was that of T.W. Schultz (1953). He noted that in the period 1910–1950 agricultural production rose by about 75 percent due to a change in inputs and in technology. The change in inputs was instigated by price change, with labor becoming more expensive and therefore replaced by machines.[14] The importance of inputs is measured by their factor shares: "Land and labor are ... very important in farming, with labor representing 46 percent and agricultural land 24 percent of all inputs used in agriculture in 1910–1914" (p. 100).

He then goes on to discuss the aggregation of inputs and to derive a measure of the overall increase in productivity by comparing the relative changes in output and input. He notices that the results are sensitive to the price weights and the period of analysis. The rise in the annual average productivity for the period as a whole with end of period prices is 1.35 percent, and with beginning of period prices is 0.8 percent.

Where does the technical change come from? Schultz (1953, p. 110) considered three hypotheses:

(1) Discoveries of new techniques are by-products of scientific curiosity and as such are unpredictable.
(2) The level of scientific activity reflects cultural and institutional values rather than the value of its fruits, and thus, the development of new techniques is not induced by market conditions.
(3) Science is supported by society because of its potential material contribution.

There is room for all three, but the gold medal is given to the last one. "Therefore, a new technique is simply a particular kind of input and the economies underlying the

[14] "Although new production techniques have been many and important, substitution among inputs is clearly evident and it is consistent with changes that have occurred in the relative prices of inputs ... labor has been withdrawn while other, cheaper inputs have been added" [Schultz (1953, p. 103)]. "United States agriculture has become increasingly dependent on inputs which are acquired from the nonfarm sector" (Ibid., p. 104).

supply and use are in principle the same as that of any other type of input. We do not wish to imply that every human activity entering into the development of new techniques can be explained wholly by considerations of cost and revenue; our belief simply is that a large part of the modern process of technological research from "pure" science to successful practice can be explained by economic analysis" [Schultz (1953, p. 110–111)]. This is the notion of induced innovation. However, "[w]e need also to explain the rate at which farmers adopt new techniques. Clearly, the mere availability of such techniques is no assurance that they will be applied in farming. The process by which farmers take on new techniques, as one would expect, is strongly motivated by economic considerations and yet very little is known about this process" (Ibid., p. 114). Although uncertainty about the new technique is important, Schultz views the new technique as a new input and suggests that the standard economic analysis be applied in the analysis of its adoption. He also recognizes the importance of credit rationing for agricultural markets. This view of technological change is related to the notion of implementation of technology discussed below.

This discussion by Schultz amplifies themes already mentioned above and puts on the agenda new ones, particularly the use of factor shares to measure the relative importance of inputs, the need to differentiate between the change in productivity due to a change in inputs and the change in technology, that the change in inputs takes place in response to changes in factor prices, and that the changes in the quality of inputs has to be taken into account in measuring factor prices. To sum up Schultz's additional observations,

O.10 Part of the change in technology is unpredictable.

O.11 Not all of what is known (in terms of technology) is actually implemented.

These are all key themes for understanding the subsequent work. To assist the discussion on the measurement of productivity, we write the production function as

$$Y(t) = F\big[A_1(t)X_1(t), \ldots, A_k(t)X_k(t), t\big], \tag{11}$$

where the A's are factor-augmenting functions or, not independently, quality indexes. Differentiate the function logarithmically, using a generic notation, $\mathrm{d}\ln x/\mathrm{d}t = \hat{x}$,

$$\begin{aligned}\widehat{Y}(t) &= \big[\omega_1(t)\big(\hat{A}_1(t) + \widehat{X}_1(t)\big) + \cdots + \omega_k(t)\big(\hat{A}_k(t) + \widehat{X}_k(t)\big)\big] + \tau(t)\\ &= [\text{aggregate input}] + \tau(t),\end{aligned} \tag{12}$$

where the ω's are weights and τ is the relative change in the total factor productivity or the 'residual'. In estimation, the A's should be included as variables in the analysis to avoid specification error.

All productivity measures are based on a comparison of changes in aggregate output with changes in aggregate input. The change in the aggregate input should measure changes in quantity that take place under constant technology. That is, the quality variables should be uncorrelated with the residual $\tau(t)$. If they are correlated, the empirical production function is a locus of points that are generated by more than one function. To illustrate, the work of children in ditch digging is not as productive as that of adults.

Therefore, adjusting the labor input by assigning different coefficients by age or gender will give a more meaningful measure of the labor input. Another example is the measure of fertilizers by their nutrient content. But most of the quality adjustments are of a different nature. A good example is the adjustment of the labor input for education where a measure of schooling multiplies the physical labor input to yield quality-adjusted labor input, measured by the total years of schooling. What is the meaning of this adjustment? If the task is digging ditches, education, at best, should not make a difference. But if there are alternatives to digging by hand, education can make a difference in the profitability of implementing these alternatives. Generalizing, an increase in the level of education, other things equal, is expected to increase the use of more advanced techniques. Thus, in this case technology is not held constant; education is a carrier of a technical change and should be treated as such. We return to this subject when we discuss the results of cross-country estimates of the production function. One implication of this distinction is that the measure of returns to scale should not include the effect of 'quality' variables that represent technology. There is no general agreement on this approach, and for alternatives see, for instance, Griliches and Jorgenson (1966).

The aggregation weights can be based on market values leading to factor shares, as done by Ruttan (1956) and Solow (1957), or by production elasticities derived from empirical production functions. Note that in the case of a Cobb–Douglas production function these elasticities are constant. Otherwise, they vary over the sample as do the factor shares, and the results vary accordingly.

Much of the work on measures of productivity change uses elasticities derived from empirical production functions. Griliches (1963a) deals directly with the effect of input quality on the measurement of productivity and, not independently, on the empirical production function. He argued for the use of the empirical production function to provide the weights for the aggregation of inputs. To this end, he fitted a Cobb–Douglas function to data for the 68 USDA regions in the US in 1949. The emphasis is on the role of education and economies of scale in accounting for productivity changes. He obtained a sum of elasticities of 1.36 from a regression without education and 1.35 with education included. Thus, the education was not the source for the sum of elasticities to exceed 1, which was taken as evidence of economies of scale. This result was incorporated in the analysis of sources of productivity growth, with the assertion that "... changes in output are attributable to changes in the quantities and *qualities* of inputs, and to *economies of scale*, rather than to 'technical change' " (Ibid., p. 332; italics by YM). "This procedure led to an almost complete accounting for the sources of output growth in the United States agriculture during 1940–60 leaving no 'unexplained' residual to be identified with unidentified 'Technical changes' " (Ibid., p. 333). The essence of that discussion is the belief that if the analysis is carried out with care, there should be no unexplained residual left.[15]

[15] This view was also repeated in [Griliches (1964)] where the empirical analysis was extended to cover 1954 and 1959. "[I]t is possible to account for all of the observed growth in agricultural output without invoking the unexplained concept of (residual) technical change" (p. 970).

There was some discomfort with the estimates, but nevertheless, those were preferred to factor shares because, relying on Schultz, the agricultural sector was perceived to be in a continuous disequilibrium.[16] As the empirical results show, education is important, the elasticities differ from factor shares, and the sum of elasticities was larger than 1. Therefore, "[t]hese findings, particularly the last two, if accepted, will account for a substantial fraction of the conventionally measured productivity increases" (Ibid., p. 336). In passing, one can question the meaning and the usefulness of the concept of equilibrium used to describe agriculture if it is thought to be in a continuous disequilibrium. Basically, it reflects an application of the concept of static equilibrium to a dynamic process. The two are not the same. We shall return to this below.

Aside from the question of the residual, can the above results be taken as indicative of economies of scale? There are two issues to be considered. First, internal economies of scale is a concept related to the cost structure of a firm and cannot be measured from regional aggregates. There are many farms of different size, and hence there is nothing in the structure of agriculture that suggests economies of scale. The optimal size depends on the technology used and the level of management of the firm. Changes in technology affect the optimal size, but this change in size is the result of the technical change. Second, there is a statistical aspect. Note that the regressions that produce a sum of elasticities larger than 1 are strictly cross-section, and hence they are subject to a bias caused by the correlation between the unobserved regional productivity level and the inputs, similar to the management bias in the analysis based on firm data. This view was taken by Kislev (1966) who analyzed data of 3,000 US districts for 1949 and 1959. To account for the unobserved regional productivity he introduced regional dummies (68 regions), and as a result the sum of elasticities declined from 1.167 to 1.05. Regional dummies do not capture the management effect, so a management bias is still present in these estimates. Very likely this is the reason that the sum of elasticities is still slightly above 1. Kislev and Peterson (1996) reexamine the evidence on economies of scale with reference to empirical results of cross-state estimates of Cobb–Douglas functions for the US[17] The sum of elasticities for each of the years 1978, 1982, and 1987 is 1.3. They do not take it as evidence of increasing returns to scale but rather as an indication of management bias. We return to this subject in the discussion of cross-country studies.

Griliches (1964) also introduces a measure for research and extension as a shifter of the production function, a practice that has been followed in other studies such as the studies based on cross-country data.

[16] In the spirit of positive economics, "[t]he most important test of the estimated production functions is not how well it fits the data it was derived from but rather whether and how well it can 'predict' and interpret subsequent behavior" [Griliches (1963a, p. 339)].

[17] The respective results for cross-state regressions for 1978, 1982, and 1987 are: land 0.1, 0.11, 0.13; labor 0.27, 0.27, 0.22; machinery 0.23, 0.27, 0.15; fertilizers and chemicals 0.27, 0.21, 0.27; and other 0.43, 0.43, 0.52.

1.4. The productivity of capital

Durable inputs are entered into the production function and in productivity analysis as stocks. This procedure is sometimes questioned (O.6), and it is suggested that the stock variable should be replaced by a flow that represents the service provided by the stock. This suggestion is based on the assumption that there is a unique variable that represents the service that can be retrieved from the analysis of annual data. Unfortunately, this is not the case. By its very nature, a durable input is purchased if the discounted expected returns from this input over its lifetime cover its cost. Thus, the service from this input is the returns over its lifetime, and this is not easily transferable to a service in a given calendar period, say a year. To sharpen the point, note that the service of a combine in the winter, when there is no harvest, is zero. However, the service for the year is positive. In some years the service is greater than in other years, depending on the area harvested and the yield, and these are affected by stochastic variables. Ex post, the value of these variables is not the same as the expected values. How are the actual values calculated? In a production function analysis, they are determined from the coefficients of the empirical equation. For instance, the coefficient of capital in a Cobb–Douglas function estimates the 'average' elasticity of capital for the sample. This can be used then to compute the marginal productivity of capital for each sample point. In some years, it may be lower than the rental cost, but this does not mean that there was too much capital in that year. The apparent overcapacity is there to provide the service in times of higher demand.

1.5. Productivity and heterogeneous technology

The foregoing discussion provides sufficient empirical evidence to evaluate the most cardinal question related to production: what is the rate, and also the nature, of technical change? Aspects of this question were addressed in one form or another in almost every empirical study of time-series data. Equation (12) characterizes much of the literature which conveys the idea that there is a unique answer to this question, and that if we work hard enough, we will find it or come close to it. Unfortunately, the matter is not that simple.

The available technology is defined as the set of all available techniques, and technical change is a change in this set. An appearance of a new technique implies a change in the available technology. In this sense, the available technology changes continuously; any new scientific publication may represent a change. However, this definition is too broad, and as such its usefulness is limited to serving as a reference point but has no operational value. The available technology contains a subset of techniques which are not implemented and thus are not observed, directly or indirectly. Therefore, there is no metric to measure the stock of the available technology or its change. Any empirical inference about technical change is based on observations and as such, by definition, is restricted to the implemented, rather than the available, technology. This is the domain of the empirical analysis.

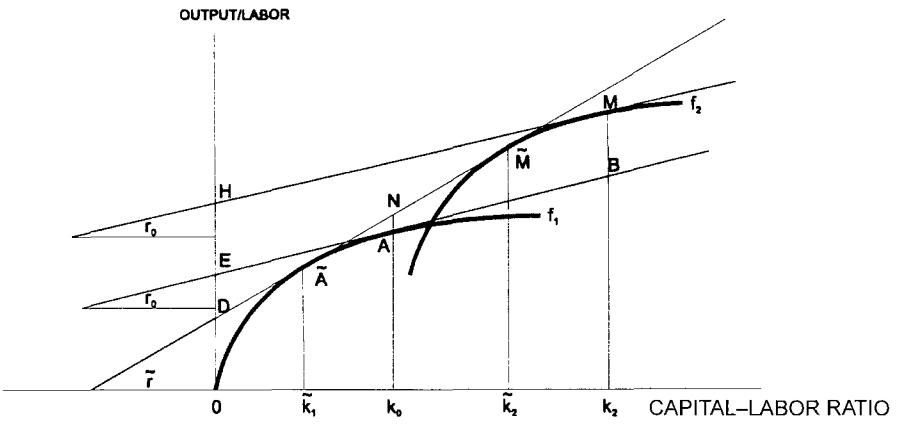

Figure 1. Resource constraint and the choice of technique.

The distinction between the available and implemented technology is not trivial if there is more than one available technique. In this case, the choice of the implemented techniques can affect the calculation of the change in the total factor productivity (TFP). To illustrate the issue, Figure 1 presents two production functions describing, say, traditional (f_1) and modern (f_2) techniques. The horizontal axis measures the input ratio, say capital-labor ratio, and correspondingly, the vertical axis measures the average labor productivity. Initially, only the traditional technique is available, and output is at point A with input ratio k_0. The response to the appearance of the modern technique may take various forms depending on the constraints to its implementation and the market conditions. If the sector is a price taker, production changes from point A to point M with input ratio of k_2. The total change in output, Y_M/Y_A, is decomposed to the input effect, Y_B/Y_A, and the relative change in the TFP, Y_M/Y_B. The point Y_B is obtained by extending the line tangent to the production function at point A to point B with capital-labor ratio k_2. If the supply of capital is initially perfectly inelastic, the input ratio remains at k_0, and resources are allocated to the two techniques to produce the output given by point N. This movement generates a relative change in TFP of Y_N/Y_A. As more capital becomes available, the movement will be along the tangent line from N to $\tilde{M}$. This movement from point N on is explained exclusively by the input change and thus shows no change in the TFP. Consequently, the resulting TFP is different from that obtained in the case of perfectly elastic factor supply. The discussion abstracts from the question of time needed to travel on each path. Actual calculations are done for data collected for calendar time, say a year. The results will differ with the changes in the pace of the yearly movement. However, when the annual results are integrated, the final outcome will depend on the path followed by the economy. Obviously, the path taken under a resource constraint will give a smaller value to the TFP. In this sense, the difference in empirical calculation of the TFP is path-dependent. The reason for the difference between the two results to the same change in the available technology is

related to the change in the factor prices, or marginal productivity. The appearance of a new technique which is both capital-intensive and more productive increases the demand for capital. When the capital supply is not perfectly elastic, its price (or its rental rate) will increase so as to internalize all or part, depending on the supply elasticity, of the technical change. Specifically, when capital is initially fixed, the subsequent movement from N to $\tilde{M}$ is fully accounted for by the change in capital availability. Thus, in the first case the contribution of the input is obtained by using the same marginal productivity in the base and new technology, whereas in the second, when the two techniques coexist, the marginal productivity of the scarce resource increases and that of the other resource declines. The resulting change of weights absorbs some of the technical change and assigns it to the inputs.

This is a remarkable result. The technical change might be of considerable magnitude and still may escape the measurement. This is the case where the bias of the technical change is in the direction of a scarce input. This applies not only to physical capital but also to human capital, and specifically to the level of education. It is in this sense that education is a carrier of technology. The literature discusses the slowdown in productivity changes in the US economy during the 1970s. Such a phenomenon is consistent with the process analyzed above where there is a change in technology but it is not captured by the calculation of productivity. The discussion is also related to adjustments in quality done in the calculation of changes in the TFP. The importance of the quality is an outcome of the technical change, and if it is considered as a contribution of the inputs, it takes away from the TFP. Thus attempts to eliminate the residual technical change by such adjustments grossly underestimate the importance of technical change (see for instance [Griliches and Jorgenson (1966)].

The implication of heterogeneous technology for empirical analysis was formulated in [Mundlak (1988, 1993)]. It is outlined in the following section. The approach was applied empirically to time series studies ([Mundlak et al. (1989)], for Argentina; [Coeymans and Mundlak (1993)], for Chile; and [Lachaal and Womack (1998)], for Canada). We will now use this framework to interpret the empirical analysis of cross-country data.

1.6. Heterogeneous technology

Let x be the vector of inputs and $F_h(x)$ be the production function associated with the hth technique, where F_h is concave and twice differentiable, and define the available technology, T, as the collection of all possible techniques, $T = \{F_h(x); h = 1, \ldots, H\}$. Firms choose the implemented techniques subject to their constraints and the environment within which they operate. We distinguish between constrained (k) and unconstrained (v) inputs, $x = (v, k)$, and assume, without a loss of generality, that the constrained inputs have no alternative cost. The optimization problem calls for a choice of the level of inputs to be assigned to technique h so as to maximize profits. To simplify the presentation, we deal with a comparative statics framework and therefore omit a

time index for the variables. The Lagrangian equation for this problem is

$$L = \sum_h p_h F_h(v_h, k_h) - \sum_h w v_h - \lambda\left(\sum_h k_h - k_0\right), \tag{13}$$
$$\text{subject to } F_h(\cdot) \in T;\ v_h \geqslant 0;\ k_h \geqslant 0,$$

where p_h is the price of the product produced by technique h, w is the price vector of the unconstrained inputs, and k_0 is the available stock of the constrained inputs. The solution is characterized by the Kuhn–Tucker necessary conditions. Let $s = (k, p, w, T)$ be the vector of state variables of this problem and write the solution as: $v_h^*(s)$, $k_h^*(s)$, $\lambda^*(s)$. The optimal inputs v_h^*, k_h^* determine the intensity at which the hth technique is implemented, where zero intensity means no implementation. The optimal output of technique h is $y_h^* = F_h(v_h^*, k_h^*)$, and the implemented technology (IT) is defined by $IT(s) = \{F_h(v_h, k_h);\ F_h(v_h^*, k_h^*) \neq 0,\ F_h \in T\}$.

The essence of the analysis is that the implemented technology is endogenous and determined jointly with the level of the unconstrained inputs conditional on the state variables. This result cannot be overemphasized, and it is essential for the interpretation of all the empirical results, regardless of specification. Of particular importance is the interpretation of the aggregate production function which expresses the aggregate of outputs, produced by a set of micro production functions, as a function of aggregate inputs. This function is not uniquely defined because the set of micro functions actually implemented, and over which the aggregation is performed, depends on the state variables and thus is endogenous. A change in the state variables causes a change in the implemented technology and in the use of inputs. It is in this sense that the function is endogenous and as such not identified. It can be identified if there are deviations from the first-order conditions. Given such deviations, we get an empirical function as $F(x, s)$. This function has a second degree approximation which looks like a Cobb–Douglas function, but where the elasticities are functions of the state variables and possibly of the inputs:

$$\ln Y = \Gamma(s) + B(s, x) \ln x + u, \tag{14}$$

where y is the value added per worker, $B(s, x)$ and $\Gamma(s)$ are the slope and intercept of the function respectively, and u is a stochastic term. This expression is given below a more descriptive structure which leads to an approach in its estimation which requires the knowledge of factor shares. The factor shares needed for this approach were not available in the cross-country application reviewed below, and therefore we do not go into it.

Variations in the state variables affect $\Gamma(s)$ and $B(s, x)$ directly as well as indirectly through their effect on inputs:

$$\partial \ln y / \partial s_h = \partial \Gamma(\cdot)/\partial s_h + \ln x\left[\partial B(\cdot)/\partial s_h\right] + B(\cdot)[\partial \ln x / \partial s_h]. \tag{15}$$

The last term shows the output response to a change in inputs under constant technology. The innovation in this formulation lies in the response of the implemented technology to

the state variables as shown by the first two terms on the right-hand side. The elasticities have a time index, which is suppressed here, indicating that they vary over the sample points. Because the state variables have a large spread across countries, the coefficients of the Cobb–Douglas function are expected to change accordingly. This is the reason for the lack of robustness in the results.

When the available technology consists of more than one technique, a change in the state variables may cause a change in the composition of techniques in addition to a change of inputs used in a given technique. In this case, the empirical function is a mixture of functions and as such may violate the concavity property of a production function. Consequently, the evaluation of empirical results should deal with the role of the state variables in production in addition to that of the inputs (or their prices in the case of dual functions). Some state variables are included in many of the studies without a reference to an explicit theory.

The state variables can be classified in the following groups: constraints, incentives, available technology, physical environment, and the political environment. There is no clear-cut separability between inputs and state variables. For instance, when capital is a constraint, its coefficient in the production function will reflect not only its productivity in a given technique but also its contribution to output through the change in the composition of the implemented techniques. A similar argument applies to the role of prices in the empirical dual functions. It is conjectured that future progress in the empirical analysis of production will have to deal more explicitly with the role of the state variables within a coherent framework. In this review, we concentrate on the role of inputs and limit our discussion of the state variables to serve this end. As such, it is incomplete but still serves a starting point to stir thinking on the subject.

1.7. Cross-country studies

The considerable spread between countries in agricultural productivity, in resource use, and in the economic and physical environment provides an important source of information for testing our understanding of the factors that determine productivity. The cross-country analysis of Bhattacharjee (1955) had no follow-up until the revival by Hayami (1969, 1970) and Hayami and Ruttan (1970). This revival added important variables that were missing in the original paper, namely measures of some capital components (livestock and machines) and of education.

The underlying assumption of these studies is that all countries use the same production function. But this assumption lacks empirical support. To get an idea of the prevailing heterogeneity, we can compare the elasticities obtained in the earlier cross-country studies (Table 1) with those obtained from country studies listed in footnote 7. For an order of magnitude, we refer to the values Hayami and Ruttan used in their exercise for sources of growth differences between countries: labor 0.4, land 0.1, livestock 0.25, fertilizers 0.15, machinery 0.1, education 0.4, and research and extension 0.15. As to the sum of elasticities, in their analysis for 1960, the estimates were in the range of 0.95–0.98. The exercise attributes about two thirds of the output differences among

Table 1
Estimated production elasticities – cross country

Study	Period	Sample	Labor	Land	All	Comments
Bhattacharjee (1955)	1948–1950	22 countries	.30	.36	1.00	
Hayami (1969)	1960	38 countries	.45[a]	.20[a]	1.00[a]	Elasticities used for productivity measures.
Hayami & Ruttan (1970)	1955, 1960, 1965	38 countries	.40[b]	.10[b]	1.00[b]	Elasticities used for productivity measures.
Nguyen (1979)	1970	40 countries	.38	.02	0.99	Regression includes education.
	1975	35 countries	.37	−.03	0.92	
Mundlak & Hellinghausen (1982)	1960–1980	58 countries[c]	.46	.16	1.00	Uses principal components method.
Antle (1983)	1965	66 countries	.33	.17	0.92	Includes infrastructure and education.
Kawagoe, Hayami, & Ruttan (1985)	1960, 1970, 1980	43 countries	.45[d]	.10[d]	1.00[d]	Elasticities used for productivity measures.

[a] Range of coefficients: Labor .43–.53, Land .18–.25, Sum 0.96–0.97.
[b] Range of coefficients: Labor .34–.49, Land .06–.12, Sum 0.94–0.98.
[c] Data is pooled for time period.
[d] Range of coefficients: Labor .41–.55, Land .01–.10, Sum 1.01–1.10.

countries to input differences and one third to differences in human capital. Subsequent studies updated and extended the analysis.

Nguyen (1979) updated Hayami and Ruttan results by computing regressions for 1970 and 1975. The results are similar to those obtained by Hayami and Ruttan with two exceptions: the elasticity of machines increased with time,[18] and the elasticity of fertilizers declined and approached zero in 1975. He finds that when education is measured as a sum of primary and secondary education, it is not significant, but secondary education alone is significant. He takes the view that the secondary education has a causal effect on productivity. Alternatively, we can interpret this result as indicative that education is endogenous, and higher productivity increases the demand for education. The adjustment to a changing economic environment is at the margin, and this places the emphasis on secondary education.

Kawagoe and Hayami (1983) and Kawagoe, Hayami, and Ruttan (1985) further update the analysis to include 1980. Like Nguyen they test for a change of coefficients over time and state that the production elasticities of conventional and nonconventional inputs remained largely the same, although some pronounced changes occurred between 1960 and 1980: the elasticity of labor declined from 0.53 to 0.41, machinery increased from 0.04 to 0.12, fertilizer increased from 0.13 to 0.25, and land increased from 0.04

[18] Similar results were obtained by Shumway, Talpaz, and Beattie (1979) for the US.

to 0.08. Thus, there is no evidence of land-saving technical change. It is hard to think of fertilizer share as being as high as 0.25, which is also in direct contrast to the results obtained by Nguyen, in which the fertilizer elasticities were approaching zero.

Another deviation from the earlier results of Hayami and Ruttan is a sum of elasticities for developing countries of about 1.3, which they take as evidence of increasing returns to scale. This magnitude affects the growth-accounting exercise because, as indicated by Equation (12), an increase in the input weights used for calculating TFP increases the contribution of the aggregate input and reduces the TFP. This explains their conclusion that the cross-country differences in output are mainly due to differences in inputs with a very small role for the residual, under 7 percent and as low as −5.5 percent. This conclusion on negligible change in the TFP is similar to that reached by Griliches (1964). As we argue below, they both are the outcome of biased coefficients which exaggerate the relative importance of the inputs. This interpretation is supported by the results reported by Kislev and Peterson (1996) who computed the Hayami–Ruttan regressions with country dummies, and the sum of elasticities declined from 1.32 to 1.077, with the latter not significantly different from 1.

A search for variables that represent the shift in the productivity level in the context of cross-country studies led Evenson and Kislev (1975) to emphasize research, and Antle (1983) to emphasize infrastructure. The problem with this group of variables is that some of them are unobservable, others are measured in some countries and not in others, and finally, because of multicollinearity, regressions do not support all of the variables that are actually used in the analysis.[19]

An implicit questioning of the assumption of uniform technology is detected in the work of Hayami and Ruttan when they divide the countries into two groups, developed and developing. This would imply that the technology changes with the level of development. However, this classification is not sufficiently informative because neither group is homogeneous. To introduce the impact of the level of development, it is more informative to include an income variable in the regression. This procedure opens up the door for extending the analysis to allow for heterogeneous technology. Mundlak and Hellinghausen (1982) remove the assumption that all countries *employ* the same production function. Instead, it is assumed that all countries have *access to the same technology* and they differ in the implementation of the technology, in line with O.11. The variables postulated to affect the choice of technology, referred to as state variables, were resource endowment and the physical environment. The resource constraint consists of physical and human capital. As no information was available on the individual components of this constraint, it is represented in the study by the per capita total output in the country. The results show a great spread in the estimates across countries and

[19] As Evenson and Kislev (1975) noted, "... with the inclusion of research variable, the fertilizer variable declines in size and significance, the same being true about the schooling coefficient .... These two variables, together with the technical education variable, served in the original Hayami and Ruttan analysis as proxies for human capital and research. These proxies are effectively replaced by genuine research variable ..." (p. 180). A somewhat similar result was obtained by Antle (1983) with an infrastructure variable.

over time which is accounted for, in part, by differences in the physical and economic environment.

All these results provide clear evidence for the lack of robustness of the empirical results, which is consistent with O.1. One possible way to stabilize the results is to choose a more flexible functional form than the Cobb–Douglas. The major changes that were introduced were the constant elasticity of substitution (CES) function by Arrow et al. (1961) and the translog function by Christensen, Jorgenson, and Lau (1973). The CES function generalized the Cobb–Douglas function by allowing a constant elasticity of substitution to differ from 1. The translog function is an example of a flexible function, a function that allows a second degree approximation to a production function. The few experiments with the CES function in agricultural economics did not prove it to be significantly different from Cobb–Douglas, and therefore it was not widely applied.[20] The situation is different with quadratic functions that have been widely used since the early 1970s, largely in connection with the dual approach, as reviewed below. From the vantage of the present discussion, we note that the main feature of a quadratic production function is to make the marginal productivities, or the production elasticities, depend on the input combination for which these coefficients are calculated. Thus, we can still postulate that all producers (or countries) use the same production function and their production elasticities vary with their choice of inputs.

Alternatively, it is possible that the producers do not use the same production function and the choice of the function is an economic decision. The variability in the state variables that exist in cross-country data offers an opportunity to gain an insight to the determinants of resource productivity. For instance, the available technology, common to all countries, varies over time. On the other hand, capital constraints and the physical environment are country specific. There are three processes which can be studied by decomposing the country-panel data to three orthogonal components to yield the regression[21]

$$y_{it} - y_{..} = (x_{it} - x_{i.} - x_{.t} + x_{..})w(it) + (x_{.t} - x_{..})b(t) + (x_{i.} - x_{..})b(i) + e_{it}, \tag{16}$$

[20] Hayami (1970) tried several modifications to the cross-country analysis. He found that a Cobb–Douglas function is not rejected when the maintained hypothesis is a CES function and that Nerlove-type distributed lags as well as serial correlation correction as suggested by Griliches gave "implausible results". Heady and Dillon (1961) discuss various functional forms used in agricultural research, including the quadratic function. Fuss, McFadden, and Mundlak (1978) discuss functional forms used in economic analysis. For an interpretation of the literature on the elasticities of substitution and their relationship to functional forms, see Mundlak (1968).

[21] Regressions that use time and country dummies provide estimates of $w(it)$, those that use only country dummies provide estimates of matrix-weighted averages of $w(it)$ and $b(t)$, those that use only time dummies provide estimates of matrix-weighted averages of $w(it)$ and $b(i)$, whereas regressions without time or country dummies provide estimates of matrix-weighted averages of all three coefficients in Equation (16). It is in this sense that the three sets of coefficients in Equation (16) constitute a canonical set.

Table 2
Cross-country panel

	Within time and country		Between time		Between country	
Variable	Estimate	t-score	Estimate	t-score	Estimate	t-score
Inputs:						
Capital	0.37	6.90	1.03	6.01	0.34	13.13
Land	0.47	3.78			−0.03	−2.82
Labor	0.08		−0.16	−0.16	0.26	13.67
Fertilizer	0.08	1.53	0.14	0.33	0.43	21.91
Technology:						
Schooling	0.09	0.55	−0.28	−0.06	0.02	0.52
Peak yield	0.83	3.80	−0.32	−0.07	0.06	4.19
Development	0.52	3.36	−0.21	−0.33	0.31	2.97
Prices:						
Relative prices	0.04	1.78	0.02	0.09	0.01	1.95
Price variability	−0.03	−0.97	−0.07	−0.26	−0.08	−2.82
Inflation	−0.00	−0.75	0.04	0.71	0.07	4.25
Environmental:						
Potential dry matter					0.16	2.68
Water availability					0.44	7.96

Note: R-square for 777 obs. = .9696, 1970–1990, 37 Countries. Source: Mundlak, Larson, and Butzer (1999).

where y is log output, x is log input (or a vector of inputs), a dot in the subscript indicates an average over the missing index, $w(it)$, $b(t)$, and $b(i)$ are the regression coefficients of the within-country-time (or, simply, within), between-time, and between-country variables respectively.

The between-time process captures the impact of changes over time in the state variables common to all countries such as changes in the available technology (technical change). The between-country process captures the impact of the country-specific variables that take place when the available technology is held constant, but other state variables differ across countries and contribute to the differences in the implemented technology. Finally, the within-country-time process represents the effect of changes in the outputs, inputs, and state variables when the available technology and the country-specific environment are held constant and thus comes closest to a production function representing what we refer to as the core technology.

This approach was used by Mundlak, Larson, and Butzer (1999) in the analysis of a sample of 37 countries for the period 1970–1990. The study differs from other studies in that it uses a new series of agricultural capital and in the state variables that were included. This choice of variables limited the sample to countries which had all the required information. We will concentrate here on the coefficients of the conventional inputs. The results are summarized in Table 2, which presents the estimated elasticities for the three regressions where the dependent variable is the log of agricultural GDP.

A striking result is the relative importance of capital. The capital elasticity is 0.37 for the core technology and 0.34 in the between-country regression. This result is quite robust to various modifications of the model and to the disaggregation of capital. On the other hand, the capital elasticity in the between-time regression is 1.03. This represents the response common to all countries in the sample. It indicates that, on average for the sample, an increase in capital was accompanied with a proportional increase in output. This strong response is consistent with the view that physical capital has been a constraint to agricultural growth. This empirical proposition is well illustrated by McGuirk and Mundlak (1991) in the context of the Green Revolution.

The between-time regression shows that the shift to more productive techniques is associated with a decline in labor. The labor coefficient in the core technology is also relatively low, whereas that of the between-country regression is more in line with the other cross-country studies. The low labor elasticity obtained for the core technology and the between-time regressions is an indication of the labor-saving technical change in agriculture, which is consistent with the slight decline of labor over time. This is not news, but it is emphasized here because it comes out of an integral view of the process which separates between the core technology and the changes that took place over time and between countries. These results highlight the importance of capital in agricultural production, an attribute critical in the understanding of agricultural development and its dependence on the economic environment. This indicates that agricultural technology is cost-capital intensive compared to nonagriculture.[22]

This last conclusion is further reinforced by the magnitude of the land elasticity in the core technology and is at variance with the view that land is not an important factor of production in modern agriculture. This view is based on an incorrect reading of the data where no distinction is made between changes in the technology and the movement along a given production function. The sum of capital and land elasticities is around 0.8 in various formulations, making it clear that agriculture should be more sensitive than nonagriculture to changes in the cost of capital, and less to changes in labor [Mundlak et al. (1989)]. This value of the sum is a bit high compared to the literature. It is possible that a different choice of countries and time periods would lead to somewhat different results. However, a sum of 0.8 for land and capital elasticities leaves room for the conclusion on the importance of capital to remain intact.

The introduction of state variables to account for technology, prices, and physical environment results in a production function that displays constant returns to scale and thus avoids the pitfalls of previous studies and the misguided conclusions that followed. Using the within elasticities from Table 2 and the median growth rates for the sample, we see that aggregate input and total factor productivity residual technical change each accounts for about one half of the total output growth of 3.82 percent per year. This evaluation of the contribution of aggregate input is substantially smaller than the rate

[22] We say that a technology is cost-capital intensive with respect to a reference technology if its factor share of capital is larger than that of the reference technology.

reported in the cross-country studies referred to above. These studies use the between-country estimates where the weight of fertilizers is high and that of land is low. The median growth rate of land in the sample was 0.12 percent and that of fertilizers was 3.04. The difference in the elasticities of these two variables accounts for much of the difference in the growth accounting. In addition, the studies that report increasing returns to scale overstate the role of inputs and understate the role of technical change.

1.8. The rate of technical change

As indicated above, all measures of technical change refer to changes in the implemented technology and thus report not only on the advances in knowledge but also on its implementation. Direct measures deal mainly with changes in the TFP and not with its bias. The latter is the subject of the studies based on duality to be discussed below. We summarize some results to give orders of magnitude to the changes in the TFP and its importance.

Ball (1985) calculates total factor productivity growth using constructed Tornqvist-Theil indexes of outputs and inputs for US agriculture for the period 1948–1979 based on data adjusted for quality variations. The inputs are labor, capital, and intermediate inputs, such as energy, agricultural chemicals, feed and seed, and miscellaneous. The result is average annual growth of productivity of 1.75 percent as compared with 1.7 percent obtained from USDA data. Capalbo and Vo (1988) review the evidence on agricultural productivity, and their result for 1950–1982 is TFP of 1.57 as compared to 1.95 as obtained by the USDA for the same period.[23] Ball et al. (1997) present the production accounts for US agriculture for the period 1948–1994 and report growth rates for the period and subperiods, based on Fisher indexes. The average growth rates for the period as a whole are 1.88, −0.07, and 1.94 percent for production (including intermediate products), aggregate input, and TFP, respectively.[24] Note that, because of the decline in the aggregate input, the growth in the TFP is larger than that in production. This result is extremely different from the studies based on cross-state data for the US, which attribute most of the change in output to inputs rather than to productivity. However, it is similar to the 1.9 percent growth result obtained by Mundlak, Larson, and Butzer (1999) for 37 countries for the period 1970–1990 discussed above.

[23] The cost shares were:

Year	Labor	Equipment & livestock	Land & structures	Chemicals	Energy	Other
1960	0.24	0.25	0.16	0.04	0.04	0.26
1980	0.11	0.21	0.41	0.06	0.04	0.17

The average annual growth rates were: output 1.76, labor −1.32, family labor −3.09, equipment 2.04, animal capital 0.38, structures and land 0.1, fertilizer 5.01, pesticides 6.07, energy 1.58, other materials 1.2, and all inputs 0.17.

[24] The change in the TFP during 1948–1979 is approximately 1.47 percent – a figure derived from Ball et al.'s (1997) results – which is lower than the figure reported in [Ball (1985)]. The difference is due to the changes in the measurement of the variables.

Jorgenson and Gollop (1992) compare the postwar productivity performance of US agriculture with sectors in the private nonfarm economy using the total price function. Productivity growth explains 82 percent of economic growth in agriculture, but only 13 percent in the private nonfarm economy. The average annual growth rate of TFP growth in agriculture during 1947–1985 was 1.58 percent, nearly four times larger than that of the rest of the economy.

Rosegrant and Evenson (1992) examine total factor productivity growth and its sources in the crops sector in India, using district panel data for the period 1956–1987. They first compute TFP and second explain its variations in terms of variables representing investments in research, extension, human capital, and infrastructure. TFP in the Indian crops sector grew during the period 1957–1985 at an average annual rate of one percent, and this accounted for about one third of total output growth in that sector. The growth rate for the same period was 0.78 in Bangladesh and 1.07 in Pakistan. Research, extension, domestic and foreign inventions, and adoption of modern varieties show statistically significant, positive impacts on TFP. The effect of the proportion of area irrigated on TFP is slightly negative, indicating that irrigation has no additional effects on productivity except through its contribution to total input levels. In any case, this procedure is only adequate if the coefficients estimated in the first stage are independent of the variables that explain the changes in the TFP. This is a strong assumption that needs empirical support, and it is inconsistent with the result reported in [McGuirk and Mundlak (1991)]. The new productive varieties are more intensive in irrigation and fertilizers, which have been scarce resources.

1.9. Primal estimates – summary

The centerpiece in primal estimation is the Cobb–Douglas function. This approach does not impose competitive conditions but instead submits them to empirical testing. Such testing often shows a difference between the factor shares and the estimated production elasticities. This is not an absolute rejection of the prevalence of the competitive conditions but rather a conditional result, based on the model used and the statistical procedure. Still it is indicative that wide gaps may exist.

Tables 1 and 2 present selected summary results of the studies reviewed as well as others with a similar message. It is noted that the elasticity of labor never exceeds 0.5, and in most cases it varies in the range of 0.25 to 0.45. This value is well below the elasticity of labor in nonagriculture.[25] If we consider all nonlabor income as capital income, the result supports the position that agriculture is cost-capital-intensive and therefore is less susceptible to increases in the wage rate than nonagriculture. Also, the labor elasticity declines with time, indicating that the technical change was labor-saving.

[25] In most studies on agriculture, output is measured as production, which includes raw materials, whereas production analysis in nonagriculture is conducted in terms of value added. Thus an exact comparison calls for applying the same output concept in both sectors. This was done in [Mundlak et al. (1989)] for Argentina, where it was found that the factor share of labor in agriculture is indeed lower than that in nonagriculture.

In country studies, the elasticity of land varies between zero in some cases to about one third. We interpret this elasticity to be a measure of the competitive position of agriculture. From the point of view of farm income it is meaningful to look at the sum of labor and land elasticities, and this sum is fluctuating around 0.5.

The sum of elasticities of farm inputs (that is, inputs decided on by the farmer, in contrast to public inputs) is used as a measure of economies of scale. In some studies based on cross-sectional data this sum is larger than 1; this was taken by the authors as evidence of increasing returns to scale. We attribute this result to statistical bias.

One justification for estimating production functions is to provide weights for the computation of technical change. However, this approach has not provided any substantive advantage as compared to the use of factor shares, even though they may not be the same as the production elasticities. The reason is that mistakes in specification and interpretation of statistical studies are often greater than the discrepancies between the factor shares and the true production elasticities. An example is the error involved in the finding of increasing returns to scale and its incorporation in the computation of total factor productivity that leads to the elimination of the residual in a comparison of growth over time or productivity differences across countries. It is tempting to speculate that such a procedure was motivated by the belief that all growth can be accounted for and therefore there should be no residual. As we take an opposite view, we do not feel that the loss of explanation involved in the reduction of the sum of elasticities to 1 causes any loss in insight; on the contrary, it directs our attention to search for an understanding of the process.

An important feature common to many of the studies is lack of robustness of the estimates and their dependence on the variables used and the sample coverage. This finding contributed to a search in three directions: 1. *Overcoming the simultaneous-equations bias caused by the endogeneity of the inputs.* As we shall see in the next section, dual estimates that were supposed to solve this problem do not produce more robust results. 2. *Algebraic form of the production function.* Indeed, the Cobb–Douglas function is restrictive, but natural generalizations enlarged, rather than shrank, the range of results. 3. *Allowing for the endogeneity of the implemented technology.* This approach utilizes the variability to improve our insight of the observed productivity differences over time and across countries.

2. The duality culture

Quadratic production functions, by their nature, contain many variables that are correlated, and therefore the estimated parameters suffer from low precision (big confidence regions) to the extent that they often do not make sense. To overcome this problem, the common procedure is to estimate the production function parameters by fitting the factor shares, with or without the constraint of the production function itself. The implicit idea is that the variations observed in the factor shares in the sample can be attributed to differences in input ratios, or said differently, to different locations on the production

function. Judging by the trend in the literature, the estimates of such functions, the most popular being the translog function, were not satisfying, and therefore a rescue was sought in the form of profit or cost functions. By this shift, the factor shares become functions of prices rather than of quantities. This shift is somewhat arbitrary in that it is not backed by any justification. We should note that the basic idea of duality is that each point on the production function corresponds uniquely to a vector of price ratios. The converse does not hold in general unless a strong assumption on the nature of the production function is imposed. Once this is imposed, then variations in prices cause, and therefore reflect, variations in quantities. This exhausts their information about technology. Hence, if regressing the shares on quantities was not satisfactory, why should prices do a better job? A plausible possibility is that the price variations cause not only movements along a given production function but also movements across production functions. This possibility is not part of the literature, but it is part of the more general framework of our discussion.

Under duality, the technology is summarized by profit, cost, or revenue functions, referred to as dual functions. The profit function is expressed in terms of factor and product prices, the cost function is expressed in terms of the factor prices and output, and the revenue function is expressed in terms of the product prices and inputs. In time-series analysis, each of these functions includes a measure of changes in technology, usually time trend. Also, the profit or cost functions are allowed to include some fixed inputs and thus are qualified as restricted or short run. Similarly, the revenue or profit functions can be restricted by the inclusion of a constraint on output (that is, a production quota).

Duality theory became a standard subject in economic analysis in the late 1960s.[26] It was adopted for empirical applications with some great hopes, but as with many innovations, the test of time has been less generous. There were several reasons for such hopes. For competitive firms, prices, unlike quantities, are exogenous and therefore when used as explanatory variables do not cause simultaneous-equations bias that is part of life in the primal estimation. This property is indeed valid but with a limited liability. First, it is not automatically applicable to data at the market or industry level. Second, it is unnecessary to estimate a dual function in order to utilize the exogeneity of prices, when this is indeed the case.

More profoundly, the econometric literature was initially motivated by the ease that duality offers to characterize the production structure.[27] Interestingly, this view paves

[26] See [McFadden (1978, p. 5) and Jorgenson (1986)] for a brief review of the history of duality.

[27] "An alternative approach to production theory is to start directly from observed economic data–supplies, demands, prices, costs, and profits. The advantage of such an attack is that the theory can be formulated in terms of causal *economic* relationships that are *presumed* [italics by YM] to hold, without intervening constructive steps required on the traditional theory. Because this approach is not bound by computational tractability in the step from production technology to economic observations, the prospect is opened for more satisfactory models of complex production problems" [Fuss and McFadden (1978, p. vii)]. Similarly "[d]emand and supply can be generated as explicit functions of relative prices without imposing the arbitrary constraints on production patterns required in the traditional methodology" [Jorgenson (1986, p. 1843)].

the way to avoid duality rather than to use it. Heuristically speaking, duality means that by following some rules (optimization), one can move from a production function to dual functions (or behavioral functions, namely product supply and factor demand) and return to the original function.[28] Thus knowing the production function, it is possible to move to the behavioral functions and vice versa. This is a simple journey under self-duality when both the technology and the dual functions have closed-form expressions. Examples are the Cobb–Douglas or the CES functions. The problem arises when self-duality does not exist, as is the case with the more complicated functional forms such as the quadratic functions. However, the move to duality in this case shifts the weight from one foot to the other in that it makes the derivation of the behavioral functions direct, but ignores the fact that questions asked about the production function itself require the exact indirect computations that were to be avoided by moving to the dual functions.[29]

For instance, given the profit, or cost, function, what is the marginal productivity of an input, and how is it affected by the input ratios? The answer to the first question is simple because by construction the competitive conditions are imposed, and therefore the marginal productivity is equal to the real factor price. The dependence of the marginal productivity on the other inputs is a question that has only a complicated answer, except when the function is self-dual. The empirical (econometric) literature on duality does not ask these questions. Thus it appears that duality is just a name, and the property is not fully exploited in the sense that the estimated behavioral functions are not used to answer questions related to relationships between inputs and outputs. However, the progress made in the ease of obtaining numerical solutions makes it possible to move from one system to the other; therefore this should cease to be an important consideration. The choice of whether to estimate a primal or a dual function should then be made on the basis of other criteria, such as statistical precision, and as argued in [Mundlak (1996a)] the dual approach to the study of the production structure is generally inferior to the direct approach. In this section we review a sample of the empirical work related to agriculture.[30]

The combination of duality and the use of quadratic functions has extended the analysis to cover topics related to the properties of the production structure and comparative statics that, with some exceptions,[31] had not been part of the agenda of most studies at the time and thereby extended the area of inquiry. Of particular interest is the attempt to fit production systems that are consistent with the assumptions of comparative statics.

[28] For a formal discussion, see [Diewert (1974)].

[29] It is therefore not surprising that a recent survey of duality contributions in production economics chooses to devote "[p]rimary attention ... to alternative ways of measuring output supply and input demand functions rather than identifying the production function" [Shumway (1995, p. 179)]. The fact is that there is little to survey on the other subjects.

[30] Shumway (1995) provides references to additional works. The survey by Jorgenson (1986) covers applications in other sectors.

[31] For instance, Mundlak (1964b) uses the second order conditions of optimization to rule out the Cobb–Douglas function as a legitimate multi-product function.

But this is done at the cost of ignoring the subjects covered in the eleven observations made above (with the exception of O.7 and O.8).

To fully describe all the properties of comparative statics, the single-output function with m inputs, or the corresponding dual function, should have at least $(m + 1)(m + 2)/2$ parameters [Hanoch (1975)]. A quadratic function that maintains the symmetry conditions has exactly this many parameters, and as such it is considered flexible in the sense that it can provide a second order approximation to the unknown true production function.[32] But since inputs tend to move together, it is statistically difficult to estimate the function directly with precision, and therefore the procedure has been to fit factor shares to the data. It is in this respect that such procedures are basically an extension of the factor share estimator.

For the dual functions to describe a production system consistent with comparative statics, they have to maintain some properties that can be tested empirically. The less trivial ones are monotonicity and convexity (or concavity, as the case may be). When the estimation is of factor demand or product supply, the monotonicity imposes signs on the first derivatives of the dual functions, whereas convexity imposes conditions on the second derivatives of the dual functions, or more to the point on the sign of the Hessian matrix. If these conditions are not met, the system is inconsistent with profit maximization. Besides these regularity properties, the dual form is used to test various hypotheses about the production structure such as separability, homotheticity, and the form of technical change.

A major shortcoming of the approach is the difficulty in achieving the regularity conditions in empirical analysis.[33] Although duality is a micro theory, many of the studies use macro data. The studies vary in functional forms used, in the type of function used in the estimation, and in the questions asked. We will try to give the flavor of these studies by sampling some that are most oriented to our needs.

[32] The parameters in question are first and second order derivatives. Their value is likely to depend on the input and output combination and thus differ with the observations. Consequently, in the event of wide variations in the sample, an approximation by a fixed coefficient function may be erroneous.

[33] In a survey of studies of US agricultural productivity, based mostly on duality, it was observed that "... empirical results and theoretical consistency are sensitive to model specification. ... Many researchers found the translog to be ill-behaved over portions of the data set, that is, monotonicity and curvature properties hold only locally [Caves and Christensen (1980)]. This was also evident in many of the models presented in this chapter. ... not all the econometric models satisfied locally the monotonicity conditions and the curvature conditions" [Capalbo (1988, p. 184)]. And in another review: "The review exposed some of the limitations of existing research. For example, it is not clear what should be done with empirical models that violate theoretical properties" [Capalbo and Vo (1988, p. 124)]. More recently, "... as most students of the existing empirical literature on agricultural supply response systems know, failure to satisfy convexity in estimated profit functions is not unique to this study" [Chambers and Pope (1994, p. 110)]. For additional supportive evidence, see also [Fox and Kivanda (1994) and Shumway (1995)].
This result had been anticipated: "Some expansions, such as the translog function ... can never except in trivial cases satisfy monotonicity or convexity conditions over the entire positive orthant" [Fuss et al. (1978, p. 234)]. This reservation is related to the functional form. However, this is not all: the major difficulty comes from the fact that the implemented technology is not constant over the sample.

Early application of duality to the study of agricultural production was made by Lau and Yotopoulos (1972) and Yotopoulos, Lau, and Lin (1976). They used a Cobb–Douglas profit function. As Cobb–Douglas is a self-dual function, it was a straightforward matter to obtain from the profit function estimates of the production function elasticities and to compare them with direct estimates of the same parameters. This comparison reveals some substantive differences. Unfortunately, such a numerical comparison of the dual and the primal estimates had no follow-up and has practically vanished from empirical analysis.

2.1. Studies based on cost functions

Define the restricted cost function

$$C(w,k,y,t)=\min_{v}\left[wv;\ y=F(v,k,t)\right], \tag{17}$$

where v is a vector of unrestricted (variable) inputs with prices denoted by w, k is a vector of constrained inputs which are assumed to have no alternative cost, y is a vector of outputs, and t is a technology index. By the envelope theorem (Shephard's Lemma)

$$\frac{\partial \ln C(w,k,y,t)}{\partial \ln w_j}\equiv S_j(w,k,y,t). \tag{18}$$

Various restrictions are imposed in empirical analysis; many of the studies assume that all inputs are unrestricted, in which case k is not part of the argument. In what follows, to simplify the notation we will use this assumption unless indicated otherwise. The empirical results depend on the structure imposed on the function. Several properties are of interest:

Homotheticity:

$$C(w,y,t)=\phi(y)C(w,t);\quad \text{Hence, } S_j(w,y,t)=S_j(w,t). \tag{19}$$

Neutral technical change:

$$C(w,y,t)=A(t)C(w,y);\quad \text{Hence, } S_j(w,y,t)=S_j(w,y). \tag{20}$$

Homotheticity and neutrality:

$$S_j(w,y,t)=S_j(w). \tag{21}$$

The cost function is expressed as a quadratic function in the variables or as a monotonic transformation of the variables, most commonly logarithmic, yielding the translog function. The share equations are then linear in the same variables. Unless indicated otherwise, the technology is represented by a time trend. The empirical analysis deals

with the estimation of the factor share equation under one of the above restrictions, often not tested empirically. There is no single central issue in these studies: different studies emphasize different topics. The most important ones are related to the behavior of factor shares with respect to changes in factor prices, the trend in the shares (time as an index of technology), and the effect of output when homotheticity is not assumed. Some studies emphasize methodological aspects by testing the properties of the function needed to describe a production system consistent with comparative statics.

Binswanger (1974) estimates a translog homothetic cost function from a cross-state data set for the US for the period 1949–1964. Agriculture is assumed to be a price-taker in all inputs, including land. He compares factor demand elasticities (evaluated at the mean) with those derived under the constraint of Cobb–Douglas. Except for land, the elasticities are near 1. They are close to the Cobb–Douglas-based elasticities for machinery and fertilizer but much lower for land (-0.34 as compared to -0.85). This result can be attributed to the fact that the model assumes a perfectly elastic supply of land, but this is not the case in reality, and the estimates reflect the data that were generated by a fairly inelastic land supply.

The cross-price derivatives of the cost function provide a measure of substitution. It is found that "[t]he best substitutes are land for fertilizer... It was a surprise... to find that machinery is a better substitute for land than for labor" [Binswanger (1974, p. 384)]. To explain the result, note that in general shocks, and specifically technical shocks, are both land-expanding and land-augmenting [Mundlak (1997)]. Technical change in agriculture caused a decline in the product price and thereby suppressed its expansion effect, so that under the new technology less land was needed to produce the demanded output. The new techniques were more fertilizer-intensive and machine-intensive, resulting in the positive association between machines and fertilizers and the negative association of these two variables with land demand.

The technical change is labor-saving and machine-using; the labor share declined at the average annual rate of 5.5 percent, and that of machines increased at the rate of 2.5 percent. Regional dummies were significantly different from zero. The inclusion of regional dummies qualifies the estimates as within-region estimates. The fact that they are significantly different from zero indicates differences in regional productivity and that the explanatory variables need not be exogenous.

Ray (1982) uses a translog cost function with two outputs, livestock and crops, in estimating the technology of US agriculture in 1939–1977. He imposes Hicks-neutral technical change and finds decreasing returns to scale for aggregate output, indicating that technology is nonhomothetic. The reason for the decreasing returns can be attributed to the fact that not all the inputs are included in the analysis, and thus, the estimates are of a short-run cost function. The average annual rate of the technical change is 1.8 percent. The own demand elasticities are less than 1. The substitution of hired labor for machines is much smaller than that between labor and fertilizers. Also, Ray finds substitution between labor and fertilizer, in contrast to Binswanger (1974), who claims complementarity.

Kako (1978) uses a translog cost function to study rice production in Japan in 1953–1970. Constant returns to scale is imposed, and technical change is measured by time trend with different slopes for three subperiods. The average percentage change in factor use during the period was: labor 2.6, machinery 3.9, fertilizers 4.4; the rice area did not change. Output grew at the rate of 2.7 percent. The input changes are decomposed to output effect, substitution (or price) effect, and technical change. The technical change was dominating for labor, whereas the output effect dominated the changes in fertilizers and machinery. Thus technical change was largely labor saving but had little effect on the other inputs. What picture does this finding portray of rice production? If the rice area did not change, it is not clear what changes in output could prompt an increase in machines. Perhaps part of the answer is related to the calculation of technical change. It is reported that 56 percent of the increase in output is attributed to technical change; thus, indirectly the use of machines is affected by technical change. We can think about these changes in terms of changes in the composition of techniques which became labor-saving and machine- and fertilizer-using. Finally, the fact that land did not change during the period is consistent with the view that land supply is far from being perfectly elastic as implicitly assumed in the formulation. As such, the results are likely to be distorted.

Kuroda (1987) estimates a translog cost function using national averages data for Japan for the period 1952–1982 and concludes that "... the production process of post-war Japanese agriculture was characterized neither by Hicks neutrality nor homotheticity. Biases ... reduced labor relative to other factor inputs ..." (p. 335).

Lopez (1980) used a generalized Leontief cost function to study the structure of production of Canadian agriculture in 1946–1977. The paper emphasizes two subjects, tests for integrability and for homotheticity. A necessary condition for integrability is symmetry of the price coefficients in the derived demand equations. Integrability is not rejected, and it is concluded that there is a production function that can represent Canadian agriculture. The idea is that the cost function can be derived from this production function. This is the idea of duality, but things are not that simple. Below we question the validity of the assumption that market prices used in an analysis of macro data are exogenous and maintain the requirements underlying the derivation of a cost function. If the assumption is violated, the estimated coefficients of the cost function would be biased. Given that the integrability conditions are met, the fitted function may be integrated to an aggregate technology, but this is not the relevant one for Canadian agriculture.[34] By way of analogy, a negatively slopped line fitted to price-quantity data need not represent a demand, or supply, function, and it may be a combination of supply and demand functions.

The factor demand equations include output and time trend. The output coefficients are significantly different from zero, indicating nonhomotheticity. The time coefficients were not significantly different from zero except for labor. This indicates neutral technical change with respect to all inputs except for labor. However, when homotheticity was

[34] On this issue see [Mundlak and Volcani (1973)].

imposed, the time coefficients became significantly different from zero, and the signs were consistent with factor-augmenting technical change. This is another illustration of the tradeoff between the inclusion of output and time trend in the equations. We discuss this finding below. The own-factor-demand elasticities are less than 1, cross-elasticities are all positive. Labor is a substitute for all inputs except for land.

Clark and Youngblood (1992) estimate a translog cost function for central Canadian agriculture (Ontario and Quebec) for 1935–1985 using a time-series approach instead of including a time trend as a technical change measure. They concur with Lopez (1980) that technical change is neutral but output is an important variable in the shares of land and fertilizers.

2.2. *What is the message?*[35]

Factor shares in agriculture have undergone changes over time; particularly, the share of labor declined, that of machinery and purchased inputs increased. How much of these changes can be attributed to economic factors? The studies reviewed above indicate that some of these changes were associated with changes in factor prices. Still, the major part of the changes is attributed to changes in output or reflects the time trend. There is a tradeoff between the role of homotheticity and neutrality of the technical change. When output was included in the equation, it tended to replace the role of the time variable.[36] This result is consistent with the fact that the new techniques are more productive and use different factor ratios than the old techniques.

Two conceptual limitations to the empirical analysis of cost functions may distort the results. First, the cost function is derived for a price-taker agent and as such does not apply to macro data where prices are determined by market supply and demand. The factor demand is derived from the cost function, and therefore it is affected by shocks affecting the cost function. These shocks are thereby translated to the factor prices. In short, factor prices need not be exogenous. This limitation applies to all studies that use market data – rather than firm data – including studies based on profit functions. This is not a trivial point because agriculture cannot be assumed to be a price-taker in the rural labor and capital markets, and definitely not in the land market.

Second, a cost function is derived conditional on output, and this is interpreted erroneously in empirical analysis to mean that output is exogenous. In general, there is no

[35] Issues related to the choice of functional form are discussed by Chalfant (1984). He argues that the translog and the Generalized Leontief cost functions are less appropriate for modeling agricultural production since they do not result in negative own-demand elasticities of substitution for all inputs. However, the estimates resulting from the use of the Fourier flexible form also failed to satisfy the negative own elasticities for all of the factors (p. 119). Lopez (1985a) discusses similar issues for profit functions.

[36] This is also consistent with the conclusion of a survey by Capalbo (1988, pp. 184–185): "Nonhomothetic functions performed better than models that maintained neutral technical change or constant returns to scale, or both". Wide variations were obtained in the level and bias of technical change, although all the reported results indicate that the technical change was labor-saving and chemical and equipment-using, whereas the results with respect to land are ambiguous.

reason to believe that the marginal cost, and therefore output, is independent of shocks to the cost function.[37] This problem is not shared by profit functions.

2.3. *Studies based on profit functions*

The profit function provides a compact form to summarize a multiproduct technology and an efficient way to introduce the properties imposed by theory on this system. This possibility is utilized in the empirical analysis, and thus there is no direct comparison with results obtained from the cost function with a single aggregate output. Also, the profit function facilitates the examination of whether the technology is that of joint production [Chambers and Just (1989)].

The restricted profit function of an individual producer is defined by

$$\pi(p, w, k, T) = \max_{y,v}(py - wv:\ y, x \in T), \tag{22}$$

where y is a vector of outputs; x is a vector of J inputs decomposed to variable, v, and fixed, k, components: $x = (v, k)$ with dimensions (J_v, J_k), $J_v + J_k = J$; T is the available technology set; p is the vector of product prices; and w is the vector of factor prices. It can be decomposed to conform to the decomposition of x. However, where ambiguity does not exist, such a decomposition is not made explicit. By the envelope theorem (Hotelling's Lemma) the product supply and factor demand functions are written:

$$y_i(p, w, k, T) = \frac{\partial \pi}{\partial p_i}, \qquad v_j(p, w, k, T) = -\frac{\partial \pi}{\partial w_j}. \tag{23}$$

The equations in (23) can be expressed also as shares. Like the cost function, the profit function is expressed as a quadratic function of a monotonic transformation of the variables. Then, Equations (23) become linear in the same variables.

Lopez (1984) estimates a Generalized Leontief profit function for Canadian agriculture, using 1971 cross-section data. The Hessian matrix (the matrix of the second partial derivatives of y_i and v_j) evaluated at the sample points has mostly the wrong sign, indicating that the profit function is not everywhere convex. The elasticities are generally low, particularly for supply (0.01 for crops and 0.472 for animal products). There is a gap between the variables used in the analysis and those assumed in the theoretical model. The paper suggests that there is sufficient variability across regions for a meaningful analysis, but this variability is in part spurious, reflecting quality variations; thus it is likely that the results reflect data problems.

Antle (1984) uses a single product translog profit function to estimate input demand and output supply functions for US agriculture for 1910–1978. Technical change is

[37] An exception in nonagriculture is the interesting study by Nerlove (1963) of the power-generating plants where the output is demand-driven and as such is exogenous.

represented by time trend and time dummies for subperiods.[38] The findings lead to the acceptance of symmetry, convexity, and structural change in the postwar period and to the rejection of homotheticity, parameter stability, and neutral technical change. Also, he finds differences in the direction of the technology bias between the pre and postwar periods.[39] Scale effects are very important post-war and are not important pre-war. "It shows that changes in factor use were more a function of technical change and a scale change in the postwar period than in the prewar period. Thus, input use in the post-war period was apparently less price responsive over time than in the pre-war period" [Antle (1984, p. 418)]. This conclusion is consistent with the world of heterogeneous technology as discussed above.

The low price elasticities are claimed to be consistent with those reported by Shumway (1983) and Weaver (1983) and as such are considered to be acceptable. This result is also consistent with many other studies of supply response reporting low supply elasticity. In our discussion of the subject at a later stage, the low elasticity is attributed to inelastic factor supply. Antle (1984) also suggests that his results are in line with induced innovations.[40] However, his argumentation indicates that the pace of the technical change was related to the implementation rather than to the pace of changes in the available technology itself.

Shumway and Alexander (1988) fit a system of five outputs and four inputs to US regional data for the period 1951–1982. They had to impose price linear-homogeneity, symmetry, and convexity.[41] It is indicated that the great variability of the results "... clearly document the importance of considering regional differences in predicting the distributional effects of potential changes in economic conditions ..." (p. 160). Technical change was not Hicks-neutral. The own-price-demand elasticities varied from 0 to -1.42, and output elasticities varied from 0.01 to 1.22, with great variations across regions.

Shumway, Saez, and Gottret (1988) estimated a quadratic profit function with five output groups and four input groups for the US for the period 1951–1982. Land and family labor are fixed; time trend represents technology. As in the previous study, symmetry, linear homogeneity, and convexity in prices had to be imposed. Estimates were obtained for regional data under the assumption that regional prices are exogenous, and for national data where the variable-factor prices were endogenized. The regional estimates are aggregated and compared with the national estimates. The output-supply and

[38] "Without time dummy variables, very small D-W statistics were obtained, suggesting misspecification" [Antle (1984, p. 417)].

[39] "The prewar is biased toward labor and mechanical technology and against land, whereas the postwar technology is biased against labor and toward machinery and chemicals" [Antle (1984, p. 420)].

[40] "Actual on-farm technology, therefore, lagged behind agricultural research, and estimates of the prewar technology should not be expected to show much evidence of technical change bias toward mechanical or chemical technology" [Antle (1984, p. 420)].

[41] "Convexity of the profit function was not maintained in the model exploration phase" [Shumway and Alexander (1988, p. 155)].

input-demand elasticities are low and become even lower when upward-slopping supply curves for the variable inputs were introduced. The low response is attributed to fixity of land and family labor.

Additional support for the proposition that techniques, outputs, and inputs are determined jointly is obtained from the fact that important properties of a production function are not maintained under aggregation over techniques: "A larger number of US parameters are significant when derived from the regional estimates (53 percent) than when directly estimated (42 percent)" [Shumway et al. (1988, p. 334)].[42] More important, "[s]ymmetry of price parameters in the system of Equations (1) and (2) was not preserved in the national aggregation" (p. 334, footnote 2). The findings also support the proposition that shocks affect land expansion and land augmentation in the same direction: "All five outputs increase as the quantity of real estates services increase .... All variable inputs are complements to real estates. Half are complements to family labor, and a third are complements to other variable inputs" (p. 334).

Huffman and Evenson (1989) fit a normalized quadratic restricted profit function with six outputs and three variable inputs to data for US cash grain farms during 1949–1974. They expand on previous duality-based studies by allowing the shares to depend on agricultural research, extension, and farmers' schooling in addition to time. The partial effect of research is in the direction of fertilizer-using and labor- and machine-saving. As research was machine-saving, the observed increased use of machines is attributed to declining prices. There is asymmetry in the explanation of the increased use of machines and the decline in the use of labor. This can be resolved by assuming that the change has been facilitated by a decline in the cost of machines and that the new machines require less labor than the old machines. This explanation is consistent with the heterogeneous technology framework. The effect of extension was small. The shadow value of private crop research is near zero, but it is high for public research. The own-price elasticities at the sample means are: fertilizer -1.2, fuel -0.72, machinery -0.61, labor -0.51, soybean 1.3, wheat 0.97, and feed grains 0.016.

Bouchet, Orden, and Norton (1989) fit a normalized quadratic profit function to data for French agriculture in 1959–1984. This was a period of strong growth, mainly in cereals, a decline in labor, and an increase in labor cost. The analysis differentiates between short- and long-run response. The supply is price responsive, but the elasticities are below 1. "However, the response to price changes are estimated to be inelastic even in the long run when usage of quasi-fixed capital and family labor have fully adjusted to optimal levels" (p. 292). The estimates of the long-run response are obtained under the implicit assumption of perfectly elastic supply of quasi-fixed inputs. When in reality the supply functions were not perfectly elastic, the estimated responses are biased downward.

[42] The standard errors for the aggregated coefficients were obtained under the assumption of independence of the regional estimates and as such are an approximation. National shocks affect all regions, and therefore their coefficients are jointly affected and thereby correlated. This may be the reason for the difference in significance levels.

The findings show that both family labor and capital have a strong positive effect on the supply of cereals, milk, and animal products. This result raises two puzzles. First, cereals is not a labor-intensive product, and therefore it is not obvious why it should have a strong positive response to changes in family labor. Second, one would expect an opposite effect of labor and capital. This similarity of effects can be explained by a strong expansion effect that dominates the substitution effect. The expansion effect is prompted by the technical change that accounts for the observed growth. Putting it all together, the observed changes can be accounted for in terms of changes in the composition of techniques.

Ball et al. (1993a) use restricted and unrestricted profit functions to evaluate the consequences of the Common Agricultural Policy (CAP). The main empirical result is that the response elasticities are low but in line with values that appear in the literature using other functional forms and less demanding models. Land and labor are taken as fixed in the evaluation, and this is the reason for obtaining low response elasticities.

2.4. Dual estimates – summary

In summarizing the foregoing findings it has to be kept in mind that the reviewed studies are mostly for the US, Canada, and Japan, so the numerical values may not be fully representative. However, the main developments in the agriculture of these countries are shared by other countries. The post-war period is characterized by a strong technical change in agriculture, both in the level and in the direction of factor use. Yields increased together with improved varieties and the use of chemicals, while labor was replaced by machines. Thus, the results have broad implications, and they facilitate the drawing of important methodological conclusions.

What distinguishes the dual approach from the primal is the appearance of prices in the empirical equation. Hence, in evaluating the performance of this approach we address the following questions:

- What has been the contribution of prices to the empirical equation?
- What additional information is obtained from the dual equations, and how can they be interpreted?
- Are the underlying assumptions of duality met?
- What are the statistical benefits of this approach?
- Where do we go from here?

The dual estimates are obtained by regressing factor shares on prices, time trend, and sometimes output. When the change in the use of inputs is decomposed to price, trend (a proxy for technology), and output effects, it is found that trend and output capture most of the changes, whereas the role of prices is the least important. Thus the contribution of prices to the explanation of inputs or output variations is rather limited.

The price elasticities of factor demand and product supply are usually obtained under the assumption that producers are price-takers in the product and factor markets. On the whole, the own-price elasticities are less than 1. There is no uniformity in the signs of the cross elasticities, but in general, most inputs appear to be substitutes. The strength of the own and cross elasticities reflects in part the fact that in reality factors' supply is not

perfectly elastic as the models assume, and therefore the results need not represent the demand-driven substitution as it is thought. This is the case with respect to elasticities related to labor, land, and capital. We further elaborate on this subject below.

With respect to other findings, interestingly, on the whole the studies based on duality do not show increasing returns to scale. Technical change, obtained by including a time trend in regressions of factor shares, is largely labor-saving, capital-using, and fertilizer-using, with the results on land being somewhat ambiguous. This is reflective of the data, which means that whatever was the effect of prices, it was not sufficient to change conclusions that could be drawn from the raw data. This does not give a strong mark to the analysis in that the results are obvious without it.

Duality between technology and prices holds under well-defined conditions that can be tested empirically. In most studies these underlying conditions are not fully met; particularly the concavity of the cost function or the convexity of the profit function is violated. Therefore, the estimated technology is inconsistent with the basic premises of the model. In a way, this is the most disappointing result because duality theory is a very powerful theory, and the question is why it does not come through in the empirical analysis. There may be more than one reason, but probably the most important one is related to the changes in technology.

One of the expected virtues of duality has been related to its solution of the simultaneous-equations bias realized in some primal estimators. However, as indicated above, in general dual estimators are inferior to primal estimators on the grounds of statistical efficiency. Where do we go from here? We return to this question at the end of the paper.

3. Multiproduct production

Most of the primal studies of production use a measure of a single output, value output, even though output consists of more than one product. The outcome is a truncated picture of the technology and limits its usefulness. Estimates based on input data aggregated over products are not sufficiently informative in that they do not provide a simple way to address questions of interest such as: What is the factor productivity in the production of a particular product? Does such productivity depend on the level of output of the other products, and if it does, is it because of overall input constraint or because of technological interdependence? Also, without a complete presentation of the multiproduct production function, it is impossible to derive the supply of the individual products. It is not due to unawareness of the importance of the complete presentation but rather due to lack of data and complexity of specification and estimation. The situation has improved considerably with the appearance of the dual approach. As the foregoing review indicates, many of the empirical studies based on the profit function facilitate the derivation of the behavioral functions, specifically product supply, without having to resort to the primal function.

The data problem is a reflection of the fact that industry statistics for agriculture do not report the inputs by products, except for land and some product-specific inputs such

as livestock. This is a convention, and by itself it reveals nothing about the nature of the production process. In principle, micro data collected from farm surveys can alleviate the problem. This is at least the case with respect to inputs which are easy to allocate to the various products, such as feeds or fertilizers, but the allocation of the use of fixed inputs requires more effort, and therefore such data are relatively scarce and do not surface with high frequency in reported studies to clarify some of the underlying issues discussed in this section.

Most farms produce more than one product, and this raises the question of the reason for the diversification. Possible reasons are:

(1) Interdependence in production where the marginal productivity of a factor of production in the production of one product depends on the level of production of another product, for example, wool and mutton, or milk and beef on dairy farms.
(2) Better utilization of some fixed inputs, or alternatively due to production quotas on some outputs, which frees resources to produce other products [Moschini (1988)].
(3) Savings due to vertical integration, where the farm produces intermediate inputs which are consumed on the farm, such as corn and hogs, or hay and livestock. Such integration saves marketing charges in the broad sense (transportation, trade margins, spoilage, etc.).
(4) Risk management.

To sort out the reasons for the diversification of production we need to go beyond the output-aggregate production function.

To put some structure to the discussion, let $T(y, x)$ denote the production set which contains all the feasible combinations of the vectors of outputs (y) and inputs (x). This set is contained in the nonnegative orthant, it is closed, convex, contains free disposals, and the origin. Its efficiency frontier, $t(y, x) = 0$, is unique. Studies with aggregate value output take the form $py = f(x)$, where py is the inner product of p and y. This is a special case of the more general presentation obtained by imposing separability on $t(y, x)$: $t(y, x) = Y(y) - X(x) = 0$ [Mundlak (1964b)]. Hall (1973) shows that this imposition is equivalent to a multiplicative decomposition of the cost function, $C(w, y) = H(y)c(w)$. The general presentation of output by $Y(y)$ has two advantages over the more restricted single aggregate output presentation: First, the function with aggregate value output is not a single-valued function and its parameters depend on the output composition along the expansion path [Mundlak (1963b)], whereas $Y(y)$ can be formulated to overcome this shortcoming. Second, it allows for interdependence in production. An application of this approach to the output aggregation of Israeli agriculture using a multi-stage CES function was made by Mundlak and Razin (1971). The limitation of this type of separability is that it is applicable only when the technology is interdependent, and the derived ratio of output prices is independent of the ratio of factor prices [Hall (1973)]. The latter, to be sure, applies also to the aggregate single-product production function.

Most agricultural production is thought to be carried out by independent techniques for individual products. In this case, the profit or the cost functions will be additive

and the supply of product j will be independent of the price of product h [Hall (1973), Lau (1978)]. We can write these functions as follows: $C(w, y) = \sum_j C_j(w, y_j)$, where $C(w, y)$ is the minimum cost of producing the output vector y at factor prices w, and similarly, $C_j(w, y_j)$ is the minimum cost of producing output y_j. A similar result applies for the profit function: $\pi(p, w) = \sum_j \pi_j(p_j, w)$. This additivity constitutes only a sufficient condition for independent production. As Shumway, Pope and Nash (1984) indicated, common constraints imposed on production may produce nonzero cross price coefficients in supply. To show this, we note that the problem under consideration is a special case of the heterogeneous technology discussed above, where the techniques are identified with the products and as such are explicit. Repeating that discussion with more details, the maximization problem is:

$$L(v_j, k_j, \lambda) = \sum_j p_j F_j(v_j, k_j) - \sum_j w v_j - \lambda\left(\sum_j k_j - k\right);$$
$$\text{subject to } F_j(\cdot) \in T;\ v_j \geqslant 0;\ k_j \geqslant 0.$$

Let F_{xj} be the vector of marginal productivity of x in the production of product j. The Kuhn–Tucker necessary conditions for a solution are

$$\begin{aligned} &(p_j F_{vj} - w)v_j = 0, && p_j F_{vj} - w \leqslant 0,\\ &(p_j F_{kj} - \lambda)k_j = 0, && pj F_{kj} - \lambda \leqslant 0,\\ &\left(\sum k_j - k\right)\lambda = 0, && \sum k_j - k \leqslant 0,\\ &v_j \geqslant 0;\ k_j \geqslant 0;\ \lambda \geqslant 0. \end{aligned}$$

Thus, even though $\partial F_{xj}/\partial y_h = 0$, it is possible that $\partial y_j/\partial p_h = \partial y_j/\partial k_j \partial k_j/\partial p_h \neq 0$, because a change in a product price may cause a change of the shadow price of the constraints in the production of that product. Therefore, when the constraints are binding, their allocation among the various products changes and causes a reshuffle of the inputs and outputs. The term joint production encompasses the two cases, interdependence in production and the sharing of constraints. The importance of the latter can be detected empirically by introducing the constraints k to the profit function.

The discussion does not indicate how to allocate the inputs to the various products. This subject is developed by Just, Zilberman, and Hochman (1983) who utilize the first order conditions for profit maximization to extract the input allocation to the individual crops. The method is further developed by Chambers and Just (1989) by introducing a flexible production function and developing a test for joint production. Without going into details, we note that in principle the allocation is determined by the Kuhn–Tucker conditions above to yield $v_j(s)$, $k_j(s)$, and $y(s)$, where $s = (p, w, k, T)$ is the vector of state variables. The two studies apply the method to the same data set and obtain plausible results in spite of the complexity in the calculations.

The essence of the discussion is that diversity in production is not necessarily a result of interdependence in production. Leathers (1991) extends the discussion to extract

implications for industrial organization. Dealing with the cost functions and taking the unconstrained cost function as the long-run function, it is implied that in the long run the constraint will not serve as a cause for diversity in production. Note, however, that agricultural production is seasonal, and since the firms are of finite size also in the long run, there is considerable scope for better utilization of resources by diversification. Just recall the old days when farm plans drawn by linear programming yielded combinations of products which utilized best the available resources.

The discussion on industrial organization deals with production at the firm level but, as we see repeatedly in this survey, we should be aware that micro theory is applied to macro data without blinking. Thus, there is another reason for diversification which is more important for the macro data – marketing costs. To put it in perspective, note that all countries produce almost all agricultural products that the physical environment permits. This can be attributed in large part to the fact that domestic production saves the various charges that are involved in international trade. Agriculture is stretched out geographically and this entails high trade costs, particularly in developing countries where the infrastructure leaves much to be desired. Finally, risk management can lead to diversification, but this is well known and need not be elaborated upon here.

4. Nonparametric methods

4.1. Description

Evidently, it is not easy to find a meaningful and robust empirical presentation of technology. The search for culprits has pointed at, among others, the parametric presentation, or functional form, of the production function, and thus the nonparametric presentation surfaced. A somewhat similar problem had been encountered much earlier in the theory of consumer choice, which sought a presentation without having to resort to the unobservable utility (objective) function. In the case of consumer choice, the empirical inference is based on the observed budget constraint, quantities, and prices. In the case of production, we observe the values of the profit (objective) function but do not observe the technology constraint, and the problem is to infer about it from the data. In the context of production, this approach was developed by Afriat (1972), Hanoch and Rothschild (1972), and Varian (1984). Recently, it has been discussed and applied to agriculture in a series of papers: Fawson and Shumway (1988), Chavas and Cox (1988, 1994), Cox and Chavas (1990), Tauer (1995), Featherstone, Moghnieh, and Goodwin (1995), Bar-Shira and Finkelshtain (1999), among others.

In describing the approach, we modify somewhat the notation used above. Let $y = (y_1, \ldots, y_H)$ be a netput vector whose positive components are outputs and the negative components are inputs, and p be the vector of corresponding prices. The profit is the inner product py. It is assumed that y comes from a feasible production set Y that maintains the free disposal property: if $y \in Y$ and $y \geqslant y'$, then $y' \in Y$.

The pivot of the analysis is the assumption that the observed netputs are optimal under the observed prices and the underlying (but unobserved) technology. Thus, if we observe

y^i and p^i, we assume that under p^i there is no netput in the production set that brings higher profit than the observed y^i. More compactly, $p^i y^i \geqslant p^i y$ for all $y \in Y$. If this holds for all the observed netputs, then it is said that the production set Y p-rationalizes the data. Then $p^i y^i \geqslant p^i y^j$ for all $i, j = 1, \ldots, n$, where n is the sample size. Varian (1984) shows that this condition guarantees the existence of a closed, convex, negative monotonic production set and referred to it as the Weak Axiom of Profit Maximization (WAPM). Bar-Shira and Finkelshtain (1999) further extend the analysis.

The underlying assumption is empirical in nature, and its validity can be tested by comparing all possible inner products between the observed netputs and prices [Fawson and Shumway (1988)]. If a netput is chosen, it should be optimal under the price regime prevailing at the time. A situation which is inconsistent with the hypothesis is when a netput is chosen even though it seems to be inferior to another netput under its own price regime: $p^j y^j < p^j y^i$ and $p^i y^j < p^i y^i$. This raises the question of why y^j was chosen in the first place when it was inferior to y^i under p^j. The negative answer is that there was a violation of profit maximization. The positive one is technical change, so that when y^j was chosen, y^i was not feasible. As technology progresses with time, we expect more recent observations to represent more productive technologies than did earlier observations. Consequently, in time series analysis, when $t > 0$, we expect $p^0 y^t - p^0 y^0 > 0$, or equivalently, $L_q = p^0 y^t / p^0 y^0 > 1$ where L_q is the Laspeyres quantity index. If this is not the case, then the conclusion is that this binary comparison is inconsistent with profit maximization.

Fawson and Shumway (1988) apply the test to regional data of US agriculture and find that the majority (typically, 80–90 percent) of the observations would be inconsistent with profit maximization if technical change were not allowed for. Featherstone, Moghnieh, and Goodwin (1995) apply the test to micro data of Kansas farms. The conditions of profit maximization, or of cost minimization, were violated by a large proportion of the observations. The number of violations declined when technical change was allowed for, but was still sizable.

When a particular netput is more profitable than another one under the two pertinent price regimes, it is concluded that it comes from a more productive technology. Based on this concept, Bar-Shira and Finkelshtain (1999) rank the technologies and apply their framework to data on US agriculture. They show that the ranking of the technologies does not always follow the chronological order, namely in some years the rank is lower than that of previous years. As no one suggests that there has been a regression in the technology of US agriculture, this finding can either be attributed to a violation of profit maximization or it may arise from more fundamental difficulties in identifying the technology through prices, an issue on which we elaborate in the discussion below. Bar-Shira and Finkelshtain quantify the technical change by computing the revenue per dollar expenses at constant prices. This is an index of change in the output-input ratio, and it is reminiscent of the early work on productivity at NBER and Schultz's (1953) discussion of productivity in agriculture. This then brings us back to square one. Finally, they examine whether the technical change is biased. Chavas and Cox (1988, 1994) go further in discussing procedures for inferring the nature of the technical change by ex-

amining what changes in the components of the netputs should be made in order to induce equality of the profits of the two netputs evaluated in terms of the base prices, or simply to bring the L_q to 1. The procedure is discussed and modified by Chalfant and Zhang (1997).

The literature deals with some more specific topics, such as separability of the technology and returns to scale. Finally, the tests discussed above are deterministic in the sense that they classify the data by those observations that are consistent with the hypothesis and those that are not. This does not take into account the possibility of errors in the data. Statistical tests have been suggested to deal with such errors. We do not cover these topics here, and we now move on to an evaluation of the method and its application.

4.2. Discussion

Under the conditions of WAPM, there exists a production set with the underlying properties needed for the production theory. Therefore, the central issue of the nonparametric analysis is to check for the empirical validity of WAPM. Note that this involves asking the same important question that was initially raised by Cobb and Douglas (1928) on the empirical validity of the competitive conditions and which received attention in the early work on the primal production function. However, as the empirical studies show, the conditions of WAPM are typically not met unless technical change is allowed for; but, to allow for technical change, the assumption that all the observations are optimal is used. At this point the common domain with the work on the primal function vanishes, and the approach becomes more similar to that of the dual function, where the optimality is imposed and not tested. This is to say that the technology is identified by the prices.

Allowing for technical change amounts to making productivity statements based on output and input indexes. It is well known that such measures are subject to the index-number bias caused by the inability to make full allowance for the substitution triggered by changes in relative prices. Thus, the method shares the problems as well as the merits of productivity measures through the use of index numbers.

It is important to note that such measures cannot differentiate between neutral and differential technical change. To show this in a simple setting, assume a cost function $C(w, A, y) = C(w_1/A_1, \ldots, w_m/A_m, y)$ where the A's are the factor-augmenting functions. Without a loss in generality, we will examine the case of a linear homogeneous production function. Also, assume $A_j^t \geqslant 1$ for all j and t. Let $A_1^t = \min_j\{A_j^t\}$ for all t, recall (19) and (20), and rewrite $C(w, A, y) = ya_1c(aw)$; $a_1 = 1/A_1$, $a_j = A_1/A_j$, $j > 1$. Thus, a_1 can be thought of as the Hicks-Neutral coefficient. Evaluate the technical change as follows, for $t > 0$:

$$\frac{C(w^0, A^0, y^0)/y^0}{C(w^t, A^t, y^t)/y^t} = \frac{a_1^0}{a_1^t}\frac{c(a^0w^0)}{c(a^tw^t)}.$$

We write more compactly $C(w^t, A^t, y^t) \equiv C(t)$, $c(a^t, w^t) \equiv c(t)$.

We now evaluate this ratio for neutral and for differential technical change. We do it under constant prices, $w^t = w^0 = w$, and for a given output, $y^t = y^0$, so that the technical change is evaluated by the savings in inputs needed to produce a given output. The inputs considered come from the input requirement set: $x^t \in V(y, t)$.

Hicks Neutral Technical Change (HNTC): Let $1 = A_1^0 < A_1^t$, $a_j^t = 1$ for all $j > 1$ and all t, hence $V(y, 0) \subseteq V(y, t)$, $C(t) = wx^t$, $wx^t < wx^0$. Imposing these conditions, we get $c(t) = c(w)$, and

$$\frac{C(0)}{C(t)} = \frac{wx^0}{wx^t} = A_1^t > 1.$$

Thus the rate of factor saving is equal to the rate of the HNTC.

Factor Augmenting Technical Change (FATC): Let $A_1^t = 1$ for all t, $a_j^t \leqslant a_j^0$ for all $j > 1$ and all t, with the inequality in effect for at least one j, hence $y^0 < y^t$ and $V(y, 0) \subseteq V(y, t)$. Impose $y^t = y^0$ and $a_j^0 = 1$, then, $wx^t < wx^0$, and the effect of the technical change under these conditions is

$$\frac{C(0)}{C(t)} = \frac{wx^0}{wx^t} = \frac{c(w)}{c(a^t w)} > 1.$$

This measure is similar to that of HNTC, but it is due to FATC; it is therefore referred to as the Neutral Equivalent of Differential Technical Change (NEDTC) [Mundlak and Razin (1969)]. The conclusion is that the ratio wx^0/wx^t is affected by neutral as well as by differential technical change, and therefore we cannot differentiate between them.

The problems in the application of the nonparametric method are similar to those faced in the applications of duality. The theory is a micro theory, and therefore its application to macro data can distort the results. Prices are not exogenous, the supply of inputs is not perfectly elastic, and in the short run, which may last for some time, there are constraints to the convergence to long-run equilibrium. We return to this topic in the discussion on dynamics below. This raises the question of how to price durable inputs in the analysis, underlining the problem that arises from the fact that the econometrician does not necessarily know the prices, or price expectations, observed by the firm and thus may use the wrong prices. All these may lead to behavior which can be incorrectly interpreted as deviations from profit maximization. To see that this can create a problem, we note that Bar-Shira and Finkelshtain (1999, Figure 8) present a graph of the profits (the product of the netput and its price) in US agriculture for the period 1945–1994. It appears that from 1958 on, with the exceptions of three years, agriculture was operating at a loss, and at times, at a big loss. During this period, output continued to increase. Thus, this suggests that somehow these prices are not the relevant prices.

All these problems occur within the traditional framework of homogeneous technology. If we allow for heterogeneous technology, additional considerations come up. First note that, by definition, the observed netputs represent the implemented technology, and as such the corresponding production sets are conditional on the state variables. As we

move from one year to the next (or across farms for that matter), the state variables may change and with them, the implied production sets. Thus, it is possible to get a regression in productivity because of the change in the underlying economic environment, as indeed it is presented in Bar-Shira and Finkelshtain (1999, Figure 9).

5. Snpply analysls[43]

5.1. Background

Analytically, the supply function of the competitive firm is the partial derivative of the profit function with respect to the product price. As we have seen above, it is one of the functions estimated in using duality to characterize the production structure. However, it has been considered as an entity by itself. The reason can be attributed to substance and history. The interest in supply analysis in agriculture had begun long before the work on the production function in agriculture and was completely disconnected from it. From its very beginning, supply response analysis was very much concerned with policy issues rather than with the application or development of formal econometric analysis. This is revealed by the titles of some of the early work: "The Farmers' Response to Price" [Bean (1929)], "The Nature of Statistical Supply Curves" [Cassels (1933)], "The Maintenance of Agricultural Production During Depression: The Explanations Reviewed" [Galbraith and Black (1938)], "Can Price Allocate Resources in American Agriculture?" [Brewster and Parsons (1946)]. Some of this discussion was motivated by the fact that agricultural production did not contract during the Great Depression of the thirties when prices of agricultural products declined substantially. The explanation for this was provided by D. Gale Johnson (1950), who indicated that not only product prices decreased in the depression, but factor prices decreased as well. This brings in the cyclical behavior of agriculture.

The central theme, the role of prices in determining output, has not changed much since. However, there are additional aspects high on the public agenda which are related to the ability to increase food supply to meet the growing demand. While the role of prices is related to the behavior under given supply conditions, the growth aspect is related to the shift in these conditions. This is a neat classification, which unfortunately does not apply to the data. Observations are determined by all the forces that affect supply, and it is therefore for the empirical analysis to sort out the role of the various factors.

Empirical supply functions regress output on prices and other variables with the purpose of extracting the output response to price. Most of the studies used aggregate time-series data, but there were some exceptions [Mundlak (1964a)]. On the whole, these studies were formulated within a static framework. As price signals do not come out

[43] In part, the discussion is based on [Mundlak (1996b)].

strong and loud in such studies, salvage is sought in using an appropriate price expectation and in a search for variables other than prices to be included in the equation.

The shift of attention to dynamic considerations gained impetus with the introduction of distributed lags to the supply analysis by Nerlove (1956, 1958). Two basic ideas are behind the formulation: adaptive expectations and partial adjustment. They both have a common outcome, a gradual adjustment in response. This is applied to expectation formation whenever a gap exists between the expected and the actual values. Similarly, it is applied to the closure of the gap between the actual output and the long-run desired output. The basic empirical equation that emerges has the form of

$$y_t = bp_t + cy_{t-1} + u_t, \tag{24}$$

where b and $b/(1-c)$ are the coefficients of short- and long-run supply response respectively. This formulation gave a neat and simple format for supply analysis and was therefore widely adapted. A summary of many studies using this framework is provided by Askari and Cummings (1976).

This efficient form for connecting the price response and the length of run has not provided the needed insight into the structure of agricultural production, nor of the origin and the nature of its dynamics [Mundlak (1966, 1967)]. In what follows we concentrate on approaches that attempt to overcome this limitation. As a background, we summarize the main empirical findings of supply analysis reported in the literature:

O.12. The short-run aggregate agricultural supply elasticity, when estimated directly, falls in the range of 0.1–0.3.

O.13. The estimated elasticities decrease with the level of aggregation. Higher values are obtained for the elasticities of individual products than for the aggregate output.

O.14. Indirect estimation of the supply elasticity, obtained through the estimation of factor demand, resulted in larger values than those obtained by direct estimation.

O.15. In the empirical analysis it was observed that adding a lagged output to a supply equation which relates output to price increases the quality of the fit and often eliminates the existing serial correlation. When measures of capital, or of fixed inputs, are added to the equation, the statistical relevance of the lagged dependent variable is reduced or vanishes. A similar result is obtained when a trend variable is added.

O.16. When the sample was divided to subperiods according to the direction of the price changes, it was found that

(a) The supply elasticity was higher for a period of increasing prices.

(b) When capital is included in the supply function, its coefficient was positive for periods of increasing prices and zero for periods of decreasing prices.

(c) When a distributed lag was used, the rate of adjustment was higher for a period of increasing prices.

O.17. The dependence of the value of the supply elasticity on the length of run reflects a constrained optimization. The severity of the constraints vanishes with time. This view leads to a formulation of a well-defined structure.

The work with duality reviewed above supplements the observations O.12 and O.13 and shows in general higher elasticities for factor demand than for the product supply which is the foundation for O.14.

5.2. Static analysis

The starting point of the analysis is the behavioral functions in Equation (23) above. The strength of the response of output and inputs to changes in prices depends on the relative importance of the restricted inputs. The unrestricted case when all inputs are variables is referred to as the long run and is represented by the following behavioral functions:

$$y^*(p, w, T), \qquad v^*(p, w, T), \qquad k^*(p, w, T). \tag{25}$$

Empirical analyses are based on dated data where some of the inputs are restricted. In this case, the response is given by Equations (23), and as such, the empirical analysis of (23) produces a restricted or short-run response. The relationship between the restricted supply and the unrestricted supply is given by the identity

$$y(p, w, k^*, T) = y^*(p, w, T). \tag{26}$$

By differentiation,

$$\varepsilon_{ii}^u = \varepsilon_{ii}^r + \sum_j \beta_{ij}^* \varepsilon_{ij}, \tag{27}$$

where $\varepsilon_{ii} = \partial \ln y_i / \partial \ln p_i$, ε_{ii}^u and ε_{ii}^r are the unrestricted (long-run) and restricted (short-run) elasticities, respectively, $\beta_{ij}^* = \partial \ln y_i / \partial \ln k_j^*$ is the production elasticity of k_j, the jth component of k, in the production of the ith product, and $\varepsilon_{ji} = \partial \ln k_j^* / \partial \ln p_i$ is the demand elasticity of k_j with respect to p_i. Thus, the long-run elasticity is the sum of the short-run elasticities and of the indirect price effect which measures the price effect on the investment in the restricted factors. The relationships in (27) are obtained under the identity in (26), and as such they are restricted to the long-run equilibrium. The demand for capital and the incorporation of nonequilibrium values in the analysis are discussed below.

It is obvious that the estimation of Equations (27) requires an elaborate statistical analysis, and we have already seen that it is difficult to get robust results. There is however a simple way to approximate meaningfully the supply elasticity. As shown in [Mundlak (1996b)], given the competitive conditions for the unrestricted inputs, the supply elasticity for a price-taker agent is approximately

$$\varepsilon = \frac{\sum_v S_v}{1 - \sum_v S_v}, \tag{28}$$

where S_v is the factor share of the vth variable input. The sum is taken over all the unrestricted inputs; it is an estimate of the scale elasticity of the 'short-run' production function, namely, the part of the function that expresses the output as a function of the unrestricted inputs conditional on the restricted ones. The scale elasticity need not be constant everywhere, as the approximation is defined locally, and thus it depends on the classification of inputs to v and k. What is important for the present discussion is that it can be evaluated in general as the sum of the factor shares of the variable inputs. This framework facilitates the derivation of orders of magnitude of the short-run supply elasticity by using empirical evidence on the elasticities of the agricultural production functions. This can be done at various levels of aggregation. To illustrate, consider the aggregate supply under the simplifying assumption that locally, the factor supply functions facing the industry are perfectly elastic and that there is no redistribution of the restricted factors among the firms in response to price variations in the short run. We assume that land, capital, and often labor are fixed in the short run. These inputs account for approximately 0.8 to 0.9 of total output, implying that the supply elasticity is between 0.11 and 0.25. The lower value is in line with the empirical results as summarized above.

The division between variable and restricted inputs is to some extent arbitrary. Such a dichotomy implies a zero supply elasticity for the restricted inputs and infinite elasticity for the variable inputs. This dichotomy is often assumed in many of the empirical analyses using derivatives of the profit function. It may hold true for the individual firm but not for the industry as a whole. Taking these considerations into account, the analysis is generalized by introducing the factor supply functions. The smaller the factor supply elasticities, the smaller the product supply elasticity [Brandow (1962), and Floyd (1965)]. Extended analytic results are given in [Mundlak (1996b)]. For instance, for a production function homogeneous of degree $\mu \leqslant 1$ in the unrestricted inputs, the supply elasticity is

$$\varepsilon = \mu\Big[(1-\mu) + \sum(\alpha_v/s_v)\Big]^{-1}, \tag{29}$$

where $s_v \neq 0$ is the supply elasticity of the vth input, and α_v is the factor share in the total cost of the variable inputs. Equation (29) generalizes Equation (28) in that when the factor supply functions are perfectly elastic for all factors, that is, $s_v = \infty$, the two equations become identical. For a linear homogeneous production function, $\mu = 1$, and Equation (29) reduces to $\varepsilon = (\sum \alpha_v/s_v)^{-1}$ which is a finite number. Thus, a constant returns to scale aggregate production function is compatible with a finite supply function because the sector is not a price-taker in some inputs.

This expression of the supply elasticity in terms of the factor shares provides the insight for the inverse relationship between the length of run and the size of the supply elasticity. The shorter the run, the more restrictions there are on factor adjustment, and therefore, the smaller the supply elasticity. Restrictions on the overall factor supply, such as farmland, do not apply to the allocation of the factor to alternative crops. For

this reason, the lower the level of aggregation of the analysis, the larger the supply elasticity (O.13).

Turning to the relationship between factor demand and the supply elasticities (O.14), we note that the price effect on input demand contains substitution and expansion effects. Of these, only the expansion effect contributes to the supply because the substitution effect of all the inputs cancels out. Technically, this is the meaning of the singularity of the Slutsky, or Hessian, matrix. This explains the findings in [Griliches (1959)] and subsequent work where the indirect supply elasticity obtained by using the factor demand elasticities gave larger values than those obtained by direct estimation of the supply function; simply, the substitution effect was not eliminated. The same holds for the estimation of the behavioral functions using the duality framework.

6. Dynamics

Equations in (23) and (25) constitute a recursive system where the long-run values of k are expressed by (25), whereas the short-run values of v and y are determined by (23) conditional on k and prices. It does not specify the time pattern of the changes in k. The analysis is now extended to deal with this subject. The extension is triggered by the fact that k affects output and cost in more than one period.

6.1. The firm's problem

It is postulated that the competitive firm chooses inputs that affect the flow of present and future profits with the objective of maximizing its expected present value. We consider here a simple case where a single output, y, is produced with a durable input, capital, k, and a nondurable, or variable, input, v, that can be hired at the ongoing wage rate, $w(t)$, using a concave and twice differentiable production function, $y = F(k, v, \tau)$, where τ represents technology. The various variables are functions of time, and the income flow at time t is $R_t = F(k_t, v_t, \tau_t) - c(I_t) - w_t v_t - q_t I_t$. Income and factor prices are measured in units of output, q and w are the real price of the investment good (I) and of the variable input, respectively, and $c(I)$ is the real cost of adjustment [Lucas (1967), Gould (1968), Treadway (1969)]. The underlying idea behind the adjustment cost is that the marginal cost of investment increases as a function of the investment rate, and hence if the firm acts too fast this cost will be excessively high. The function is convex in I (or in the ratio I/k). Let r be the interest rate, $\beta = (1+r)^{-1}$ is the discount factor; the optimization problem calls for selecting the time path of inputs $\{v_j, k_j\}$ that maximizes the expected value of the firm at the base period, 0,

$$\max_{k_{j+1}, v_j} \left\{ \mathrm{E}_0 \left[\sum_{j=0}^{\infty} \beta^j \left[F_j(k_j, v_j, \tau_j) - w_j v_j - q_j I_j - c(I_j) \right] \right] \right\} \tag{30}$$

subject to $I_j = k_{j+1} - (1-\delta)k_j$, the initial value k_0, and terminal conditions, where k_j is the capital stock at the beginning of period j, and δ is the depreciation rate. The expectation, E_0, is taken over the future prices and the technology whose distribution is assumed known.[44]

To obtain the first order conditions we first differentiate (30) with respect to the non-durable inputs, v_j, to obtain:

$$E\left[\frac{\partial F(\cdot)}{\partial v_j} - w_j\right] = 0. \tag{31}$$

By assumption, the input v_j at any time j has no effect on the revenue in subsequent periods, and therefore its level is determined by equating the expected value of the marginal productivity to that of its real price in each period, as shown by Equation (31). Consequently, the optimization problem can be solved in steps. First, determine for each period the optimal level v_j as a function of prices and k_j, and substitute the result in the production function to obtain the function, $F(k_j, s_j)$, where $s_j \equiv (\tau_j, w_j, q_j, r, \delta, c)$ is the vector of the exogenous variables. The second stage consists of solving

$$\max_{k_{j+1}}\left\{E_0 \sum_{j=0}^{\infty} \beta^j \left[F_j(k_j, s_j) - c(I_j) - q_j I_j\right]\right\} \tag{32}$$

subject to $I_j = k_{j+1} - (1-\delta)k_j$.

Label the rate of capital appreciation $\hat{q} \equiv \dot{q}/q$ and $\tilde{q}_j \equiv q_j[r + \delta - (1-\delta)\hat{q}_j]$, which is the rental cost of capital, or briefly the rental rate, evaluated at time j. It is the product of the initial price of the capital good, q, and the annual "charges" consisting of the discount and depreciation rates, adjusted for the expected capital gain, $\hat{q}$. Similarly, $\tilde{c}_I \equiv c_I(j)[r + \delta - (1-\delta)\hat{c}_j]$ gives the change in the adjustment cost due to a change of the timing of a unit of investment, *on the optimal path*, from one year to the next. Differentiate (32) with respect to k_{j+1} and rearrange the result to obtain, for the case when an internal solution exists,

$$E_0\left\{\beta F_k(j+1) - \left[\tilde{c}_I(j) + \tilde{q}_j\right]\right\} = 0, \tag{33}$$

where we use the notation $F(k_j, s_j) \equiv F(j)$ and similarly for other functions, and the subscripts k and I indicate the direction of the partial derivatives of the functions in question. Under static expectations, where the present prices are expected to remain constant indefinitely, $E(\hat{q}) = E(\hat{c}) = 0$, and (33) becomes $\{\beta F_k(j+1) - (r+\delta)[c_I(j) + q_j]\} = 0$. In the absence of adjustment cost, this condition reduces to the equality of the marginal productivity of capital and the rental rate [Jorgenson (1967)]. This condition applies to every point on the optimal path. The addition of the adjustment cost affects

[44] The terminal condition is $\lim_{j\to\infty} E_0\{\beta^j[F_k(j) - c_I(j) - q_j]k_j\} = 0$.

the rental rate, and thus it affects not only the pace of investment but also the optimal level of capital.

The solution can be expressed in terms of the shadow price of capital defined as the present value of the marginal productivity of capital, net of the adjustment cost, in present and future production: $S_t \equiv \sum_{j=0}^{\infty} h^j F_k(t+j)$, where $h \equiv (1-\delta)\beta < 1$. The system can be solved to yield

$$\mathrm{E}_t\{S_t - (q_t + c_I(t))\} = 0. \tag{34}$$

This condition states that investment is carried out to the point where the shadow price of capital generated by the investment is equal to the cost of investment including the cost of adjustment. The marginal productivity depends on the technology and the inputs at the various points in time, and therefore its evaluation requires an assumption that the investment under consideration is the only investment to be made. If other investments are contemplated, the marginal productivity would have to be evaluated conditional on such investments.

6.2. *Discussion*

The condition in (31) is extremely important for empirical analysis in that it implies that along the optimal path, the use of the inputs which have no effect on the revenue or the cost in subsequent periods is determined by equating the marginal productivities to their real prices in each period. This leads to a recursive system [Mundlak (1967)]. First, we determine for each period the optimal levels of the variable inputs as functions of the exogenous variables, including prices and $k(t)$. Second, we solve for $k(t)$ on the optimal time path:

$$k^*[\mathrm{E}(q, \hat{q}, \delta, r, c, w, p, T)], \tag{35}$$

where we insert p, the product price, explicitly. All the variables in (35) are functions of time. The introduction of the intertemporal optimization results in replacing $k^*(\cdot)$ in (25) with (35), thereby adding exogenous variables as well as uncertainty with respect to the future time path of the exogenous variables. However, the recursive structure remains the same.

6.3. *The role of prices and technology*

The solution is quite sensitive to changes in the exogenous variables. To gain some insight into the meaning of the solution, we use a Cobb–Douglas production function, $y = Av^a k^b$. The first order condition in (31) provides a solution $v = (a/w)y$ for the nondurable input. This solution is substituted in the production function to yield, with some simplification,

$$Y = (Aa^a)^{1/(1-a)} w^{-a/(1-a)} k^{b/(1-a)}. \tag{36}$$

The marginal productivity of capital conditional on w is[45]

$$\left.\frac{\partial y}{\partial k}\right|_{w} = \frac{b}{1-a}\left(Aa^{a}\right)^{1/(1-a)} w^{-a/(1-a)} k^{(b+a-1)/(1-a)}. \tag{37}$$

This derivative is equated to the rental price of capital to provide a solution for k^*, when such a solution exists.

Equation (36) is the short-run supply function conditional on k. Output declines with w, but as w is the ratio of nominal wage to output price, p, output increases with p. To simplify the discussion without a loss in generality, we continue by ignoring the adjustment cost. The condition in Equation (33) simplifies to

$$\mathrm{E}_0\left\{\beta F_k(j+1) - \tilde{q}_j\right\} = 0. \tag{38}$$

The long-run values (starred) are obtained by using Equations (36) and (38) to yield

$$k^* = (b/\tilde{q})y^*, \quad y^* = \left(Aa^a b^b\right)^{\varepsilon} w^{-a\varepsilon} \tilde{q}^{-b\varepsilon}, \quad \varepsilon = 1/(1-a-b). \tag{39}$$

Prices affect the desired capital directly through the rental rate and indirectly through the effect on the optimal output. It is important to differentiate between the direct and the indirect price effect. A change in the wage rate has only an indirect effect on capital with an elasticity $E_{k/w} = -a\varepsilon$. The elasticities of the real rental rate, $E_{k/\tilde{q}}$, are -1, $-b\varepsilon$, and $(a-1)\varepsilon$ for the direct, indirect, and total effect respectively. Similarly, the elasticities of capital with respect to a change in the product price are 1, $(a+b)\varepsilon$, and ε for the direct, indirect, and total effect respectively. Note that the indirect effect $(a+b)\varepsilon$ is considerably stronger than the direct effect. It is useful to illustrate the order of magnitude of the elasticities in question for arbitrary values of the parameters (Table 3). The elasticity of labor is maintained at 0.3 for the three cases, whereas the elasticity of capital varies from 0.6, a highly capital-intensive process, to 0.1. Note that 0.1 is approximately the estimated elasticity of machinery in many studies, whereas a value of 0.3 represents a broader capital aggregate, including structures. The difference $1-a-b$ is the share of fixed factors which vary across cases. In the first case it would be management, whereas in the last case it might also include land. The values in this table provide an insight into the interpretation of the empirical results.

To simplify the discussion, we have abstracted from taxes. To add taxes, they have to be inserted in the income expression in (1), and the prices in the foregoing results would have to be adjusted for taxes [Jorgenson (1963)]. The empirical evaluation of the effect of taxes is done in two steps: first, evaluate the effect of the tax on the time path of the rental rate; and second, determine the response of investment to price. It is the latter that is the focus of the empirical analysis.

[45] This derivative is evaluated for v kept at its short-run optimal level, which is different from the derivative conditional on v derived from the production function: $\frac{\partial y}{\partial k}|_{v} = b\frac{y}{k}$.

Table 3
Capital-demand elasticities

	$a=0.3, b=0.6, \varepsilon=10$			$a=0.3, b=0.3, \varepsilon=2.5$			$a=0.3, b=0.1, \varepsilon=1.67$		
Prices	*D*	*I*	*T*	*D*	*I*	*T*	*D*	*I*	*T*
W	0	−3	−3	0	−0.75	−0.75	0	−0.5	−0.5
$\tilde{q}$	−1	−6	−7	−1	−0.75	−1.75	−1	−0.17	−1.17
P	1	9	10	1	1.5	2.5	1	0.67	1.67
NTC	0	10	10	0	2.5	2.5	0	1.67	1.67

Legend: D = Direct, I = Indirect, T = Total, W = wage rate, P = product price, $\tilde{q}$ = rental rate, NTC = Neutral technical change.

Neutral technical change is perceived as a change in the multiplicative coefficient (A) of the production function. It affects output and thereby the desired capital level without affecting the capital-output ratio. The demand elasticity with respect to neutral technical change is equal to ε. Capital-using technical change, captured here as an increase in b, generates an increase in capital demand and in the capital-output ratio. The overall effect of such a technical change on output depends on what happens to the degree of the function. When the degree is held constant, an increase in b implies a decline of a, and therefore, without imposing a more detailed structure, the net effect on output is ambiguous.

To summarize, the expected magnitude of the estimated demand elasticities depends strongly on what variables are held constant in the sample, and therefore we can expect a considerable variability in the empirical results.

6.4. Disinvestment

In general, empirical analysis treats positive and negative accumulation symmetrically even though the costs involved are completely different. The cost of acquisition of a new tractor is different from the selling price of a used one. Implications of this additional detail are discussed by Glenn Johnson (1958), Edwards (1959), Johnson and Quance (1972, pp. 185–195), and more recently by Chavas (1994) and Hamermesh and Pfann (1996). To place this detail in perspective, we note that on the whole, agricultural investment is positive for most of the time, and therefore the subject of disinvestment is of secondary importance and does not affect our views on the development of agriculture. Its empirical importance is largely limited to the analysis of cyclical behavior and the analysis based on micro data which include firms with zero or negative investment.

There are several important reasons for the difference between the acquisition and the selling price. First, the service life of the new capital good is longer than that of the used one, and therefore it is more valuable. Conceptually, this aspect can be incorporated into the analysis by disaggregating the capital goods by age and vintage and pricing the different goods accordingly. The optimization problem of the price-taker farmer would

then include acquisition prices by age and vintage instead of one price. If an old machine is sold, someone is buying it because it meets his needs. This indicates that there is a market for all types of machines which are actually traded. The extension of the analysis to include this kind of heterogeneity should give qualitatively different results from the one obtained when the farmer is restricted from purchasing the used equipment (who will then buy it?), as the standard model assumes. The interesting question is what the qualitative effect is.

Second, part of the gap between the price of new and used equipment can be attributed to marketing charges and asymmetric information of the pertinent agents. Third, there is the cyclical element. There is a tendency to sell unutilized capacity in bad times when the excess demand for capital goods is declining and with it the price of the used equipment. The cyclical price behavior is likely to differ according to the origin of the capital goods. Used machines are supplied by farmers, and for our purpose they are expected to behave as do capital goods of agricultural origin. Their price is determined endogenously within agriculture and reflects the expected stream of the marginal productivity of capital over its remaining lifetime. To trace the consequences of this extension, it is necessary to work out the market equilibrium for used equipment. This will result in a market clearing price, and used equipment will be employed according to conditions analogous to Equation (34). New machines are of nonagricultural origin, and their supply price reflects the conditions in nonagriculture. Therefore the price may be less sensitive to the cyclical conditions in agriculture as compared to used machines. To sum up, the introduction of a second-hand market adds details to the analysis but not a new theory.

The asymmetry between investment and disinvestment is more pronounced in models with internal adjustment costs. Obviously, a demolition of a building or a slaughter of a cow does not stretch out over time. The symmetry assumption simplifies the formulation, but it is unrealistic. Its restrictive nature goes undetected because much of the empirical work is based on aggregate data. However, there are some exceptions such as Chang and Stefanou (1988) and Lansink and Stefanou (1997).

6.5. Empirical investment analysis

In general, time series of aggregate investment show a positive serial correlation. The determination of the source for this dynamic relation is a key question in investment research. There are two basic approaches. Initially, the dynamics was superimposed on the model, and we therefore refer to it as exogenous dynamics. Alternatively, the dynamics can be developed from the theory, such as in the case of models based on adjustment cost, and it is therefore referred to as endogenous dynamics.

Aside from the pattern of the dynamics, the empirical analysis should reveal the response of k^* to changes in its determinants, where k^* is unobserved and therefore is replaced by the actual capital stock, or changes in it. The actual capital stock by itself is not a well-defined variable, but in this discussion we will ignore the issues involved in the construction of the capital stock.

6.6. Exogenous dynamics

For a variety of reasons, there is a time difference between the date of a firm's decision on a new investment and its completion. The implication is that a decision taken by the firm in a given year may affect investment in future years, or alternatively, the investment in a given year reflects past decisions and, more so, past signals. Such a time distribution of the response was a major justification for the distributed lags analysis, referred to as the flexible accelerator models, introduced by Chenery (1952) and Koyck (1954). In such models, the actual capital stock differs from the desired stock. Koyck's formulation uses geometric weights to express the current capital stock as a weighted average of past values of desired capital. This process can be presented by an adjustment equation

$$k_t - k_{t-1} = \mu\left(k_t^* - k_{t-1}\right), \tag{40}$$

where μ, $0 \leqslant \mu \leqslant 1$, is the coefficient of adjustment. Nadiri and Rosen (1969) extended this model to more than one quasi-fixed factor.

The desired capital is unobserved. In the case of a Cobb–Douglas production function, the desired capital stock is proportional to the long-run output, and the latter can replace the first. Introducing this substitution into Equation (40) and simplifying, we can write the following investment function, where I_t is the *net* investment in year t,

$$I_t = \mu\gamma_0 + \mu\gamma y_t^* - \mu k_{t-1} + \text{error}. \tag{41}$$

However, the replacement of k^* by y^* is of little help because the latter is also unobservable. In practice, actual output is used instead in empirical analysis [Jorgenson (1963)]. In so doing, the difference between the short- and the long-run supply is overlooked. The elasticities for long-run response express the response with respect to lasting price and technology changes. Transitory price changes are likely to affect output according to the short-run supply function, but as such should not affect the capital demand. Consequently, the variable used in the analysis measures with error the relevant variable and thereby introduces a downward bias in the estimation [Mundlak (1966)].[46] The problem can be overcome by aggregating the variables over time and thereby reducing to a large extent the effect of the transitory variations [Mundlak (1964a, Chapter 6)].

The underlying assumption in Equation (40) is that the adjustment of the actual stock to changes in the desired stock is gradual, but this is not always the case. Often, there are distinct scale economies in the size of the investment, where the unit cost declines with the size of the project, and the optimal size of the investment unit exceeds the demand or requires more resources than are currently available. Consequently, the firm may delay the investment until it is justified to construct a larger project at a lower unit cost (Ibid.).

[46] For more detailed discussion of this subject, see [Mundlak (1964a)].

The phenomenon of lumpy investment at intervals longer than a year is inconsistent with the adjustment cost assumption. However, this is not detected in empirical analysis which uses macro data obtained as aggregates over firms and as such conceal it. Again, with micro data the problem can be overcome by aggregating the variables over time and thereby reducing the importance of the exact timing of the investment (Ibid.). This problem has resurfaced in the context of analysis based on adjustment costs, and we return to it below.

6.7. Endogenous dynamics – the primal approach

There has been a great deal of empirical work based on the Euler equation on nonagricultural data. The equation involves unobservable variables, and to overcome this limitation, alternative approaches have been taken; these are reviewed by Chirinko (1993) and Galeotti (1996). To illustrate the basic issues at stake, we present an empirical version of Equation (33), with the assumption that $c(\cdot) = (c/2)I^2$ so that $c(\cdot)$ does not depend on the capital stock. Let z be the expected gap between the marginal productivity of capital and the rental rate, $z_{t+j} \equiv \mathrm{E}_t\{\beta F_k(t+j+1) - \tilde{q}_{t+j}\}$. Rearranging Equation (33) subject to the assumption on the adjustment cost, it follows that

$$\mathrm{E}_t(I_{t+j}) - h\,\mathrm{E}_t(I_{t+j+1}) = \frac{1}{c} z_{t+j}. \tag{42}$$

An expected decline in the rental rate or an expected increase in the productivity of capital causes an increase in z, and hence the difference between current investment and expected next-year investment increases. This means that at the margin, current investment increases in order to take advantage of the current opportunities.

For the purpose of estimation, F_k is spelled out explicitly in terms of its arguments, and thus the parameters of the production function enter the equation. Similarly, in some applications, the cost of adjustment is formulated so as to depend on some variables, including output. When the marginal productivity of capital and the adjustment costs are written explicitly in terms of their determinants, the empirical equation contains output and prices. The empirical equation is then used to estimate the parameters of the production function, of the adjustment-cost function, and of h. Unlike in the exogenous dynamic models, it is assumed here that the observed capital stock is *always* equal to the optimal one.

There are several problems in using this equation for empirical analysis. First, in this formulation the adjustment-cost parameters are, by assumption, the only source for the dynamics. When in reality the time pattern of investment is affected by other causes, their influence will be captured by the cost of adjustment parameters, and the empirical analysis will give a distorted picture of the dynamics. Second, the Euler equation, (42), provides arbitrage conditions between adjacent periods which have to be met on the *optimal path*. When the observations are located off the path, this condition is inconsistent with the data. If the model is stable, deviations from the optimal path generate a

correction toward the path. This correction is not described by the model, but it is empirically important and as such it affects the estimates. This may be the reason for the fact that empirical estimates obtained from the Euler equations do not produce robust results. Third, the Euler equation is not an efficient way to estimate the parameters of the production function. As argued earlier, it is more efficient to estimate the production function directly. Fourth, recall that $h = (1 - \delta)/(1 + r)$, so that h is not a stable parameter and should be treated as a variable. When h is treated like a constant, variations in h are captured by the equation error, and as such the error is not independent of the investment term on the right-hand side of the equation. This causes a bias in the estimate.

6.8. Endogenous dynamics – the dual approach

The dual approach, as developed by McLaren and Cooper (1980) and Epstein (1981), has provided an elegant framework to deal simultaneously with several issues of dynamic adjustment in a practical fashion. It has been applied in agricultural economics research, reviewed below, and it is therefore summarized here.

Following the literature, the presentation is in terms of continuous time, and the cost of adjustment appears as an argument in the production function. A crucial element in this framework is the assumption of static price expectation whereby the present prices and technology are assumed to remain constant indefinitely. Modifications of this assumption are discussed below.

The production function, $F(k, I)$, is expressed in terms of the quasi-fixed factors, k, and the investment, I.[47] The variables are vectors of comparable dimensions. A partial list of the regularity conditions on the production function includes: $F_k(\cdot) > 0$, $F_I(\cdot) < 0$, and $F(\cdot)$ is strongly concave in I. The optimization calls for:

$$J(s) = \max_I \int_0^\infty e^{-rt}\left[F(k, I) - \tilde{q}'k + J_{k'}(I - \delta k)\right] dt \tag{43}$$

subject to $k(0) = k_0$, and the terminal conditions. $J(s)$ is the value function, a prime means transpose, $\tilde{q}$ is the vector of rental rates, $s = (k, \tilde{q}, r, \delta)$ is the vector of exogenous variables, J_k is the vector of multipliers of the constraint $\dot{k} = I - \delta k$, and as such it represents the shadow price of capital. Note that (43) is expressed in terms of the rental rate, unlike the argument of (30), which is expressed in terms of the price of the capital good. Also, under static expectations, $\tilde{q}$ does not contain the capital-appreciation term. This difference in formulation can be of significance in the case of nonstatic expectations. In what follows, unless indicated otherwise, r and δ are assumed to be constant. All the variables are functions of time and, unless needed, the time notation is avoided.

[47] Initially, all inputs can be considered to be quasi-fixed, and it is up to the analysis to determine if a particular input is variable. Alternatively, the production function can be the concentrated function in the quasi-fixed variables.

Because the prices and the technology are assumed constant, only their current values matter. This is the major analytic payoff of the assumption of static expectations. Consequently, the problem becomes similar to that of the duality used in the static analysis. The difference between the two models is in the nature of the solution; in the dynamic case, it consists of the time path of the control variables.

Under the regularity conditions on F, the value function J satisfies the Hamilton–Jacobi–Bellman equation [Kamien and Schwartz (1991, p. 261)]:

$$rJ(s) = \max_{I}\{F(k, I) - \tilde{q}'k + J_k(s)'(I - \delta k)\}. \tag{44}$$

A partial list of the regularity conditions on the value function includes: $(\delta + r)J_k + \tilde{q} - J_{kk}\dot{k} > 0$ (equivalent to $F_k > 0$), $J_k > 0$ (positive shadow price of capital; follows from the adjustment cost assumption of $F_I < 0$), and a necessary condition that J is convex in prices (because J is a maximum problem).

The behavioral functions are derived by differentiating $J(\cdot)$ with respect to the exogenous variables to yield a generalized Hotelling's Lemma. Specifically, a differentiation with respect to $\tilde{q}$ and rearrangement yields:

$$\dot{k}^* = J_{\tilde{q}k}^{-1}\big(rJ_{\tilde{q}}(k, s) + k\big), \tag{45}$$

where we write $J(k, s)$ to remind us that k is an argument of J. Thus, the following holds on the optimal path:

$$rJ(s) \equiv F(k, \dot{k}^* + \delta k) - \tilde{q}'k + J_k'(s)\dot{k}^*. \tag{46}$$

The steady state value of k is obtained by setting $\dot{k}^* = 0$ and solving:

$$k^* + rJ_{\tilde{q}}(k^*, s) = 0. \tag{47}$$

Given the regularity conditions on $J(\cdot)$, a duality between $F(\cdot)$ and $J(\cdot)$ is established. Let

$$F^*(k, I) = \min_{\tilde{q}}\{rJ(k, \tilde{q}) + \tilde{q}'k - J_k(k, \tilde{q})'(I - \delta k)\}. \tag{48}$$

Heuristically, the duality prevails if J derived from (43) is used in (48) to derive $F^*(\cdot)$, and $F^*(\cdot) = F(\cdot)$. Inversely, if F derived from (48), by using J that maintains the regular conditions on J, is used in (43) to derive J^*, then $J = J^*$. This is the meaning of the duality, but as in the static case, this relation is seldom exploited in empirical work. However, there is a revealed difference in aspiration between the static and dynamic analyses. As discussed above, the empirical duality analysis sprung up as an alternative to the primal approach for estimating production functions. The dynamic analysis is focused on the derivation of the demand for the quasi-fixed factors of production. As

such, the interest is in the empirical performance of (44) and (47) and the conditions underlying their derivation.

The empirical implementation requires algebraic formulation of the value function. The quadratic function, in the pertinent variables (or a monotone transformation thereof, such as logarithms or power functions), has been widely used because of its convenience:

$$J(s) = a_0 + \left(a_k' a_{\tilde{q}}'\right)\begin{pmatrix} k \\ \tilde{q} \end{pmatrix} + \frac{1}{2}\left(k'\tilde{q}'\right)\begin{pmatrix} A_{kk} & A_{k\tilde{q}} \\ A_{\tilde{q}k} & A_{\tilde{q}\tilde{q}} \end{pmatrix}\begin{pmatrix} k \\ \tilde{q} \end{pmatrix}, \tag{49}$$

where a_k, $a_{\tilde{q}}$, k, and $\tilde{q}$ are column vectors, and the A_{ij} are matrices of conforming dimensions. Given (49),

$$J_{\tilde{q}} = a_{\tilde{q}} + A_{\tilde{q}k}k + A_{\tilde{q}\tilde{q}k}\tilde{q}; \quad J_{\tilde{q}k} = A_{\tilde{q}k}. \tag{50}$$

Substitute in (47) and impose $k = k^*$:

$$k^* = -r(I + rA_{\tilde{q}K})^{-1}(a_{\tilde{q}} + A_{\tilde{q}\tilde{q}}\tilde{q}), \tag{51}$$

where I is the identity matrix. Substitute (50) in (45) and simplify using (51),

$$\dot{k}^* = M(k - k^*); \quad M \equiv (I + rA_{\tilde{q}k})^{-1}, \tag{52}$$

where M is the adjustment matrix. Note the similarity of Equations (52) and (40). In addition to the fact that (52) is a differential equation and (40) is a difference equation, Equation (52) is a multivariate equation and M is expressed in terms of coefficients of the value function. Otherwise, in empirical applications, the two versions are similar in form, so that the foregoing discussion provides a foundation for the distributed lag formulation. Using a discrete time approximation, the empirical equation can be written as

$$k_t = (I - M)k_{t-1} - Mk_t^*. \tag{53}$$

The adjustment matrix, M, is constant, but under a different specification of the value function it can become a function of some exogenous variables.

6.9. Empirical investment analysis in agriculture

The following review of individual studies is intended to span the space of the empirical parameters, and to convey the cumulative experience which should help us in forming a view of the scope of the various approaches and to learn from their inherent difficulties. This should help in outlining the strategy for future research. Our discussion is limited to the estimation of investment functions and will skip over the important conceptual

and practical issues involved in measurements of capital (see for instance: [Griliches (1963c), Ball et al. (1993b), Larson et al. (1999)].

Unlike studies of production or supply functions, there are only a few empirical studies of investment in agriculture using the direct or primal approach. Griliches (1960, 1963c) studied the demand for tractors in the United States in 1921–1957 using a distributed lag framework where the desired stock is determined by the real price of tractors and by the interest rate. The results show the importance of price variables as determinants of investment.

Heady and Tweeten (1963, Chapter 11) analyzed the purchases of all farm machinery in the United States in the period 1926–1959, excluding 1942–1947. They report a garden variety of regressions. The core explanatory variables are machines-to-commodity price ratio, a ratio of equity to liabilities of the farm sector, or alternatively a measure of farm income, a time trend, and in some cases, the lagged value of the dependent variable. They conclude that "... a 1 percent increase in the price of either trucks, tractors or equipment aggregate ... is predicted to increase respective annual purchases 1 percent; stock 0.2 percent in one or two years. In four years the elasticity of machinery purchases Q_i with respect to P_i remains about unity, but with respect to P_R [commodity price – YM] is 2 or more. A sustained 1 percent rise in prices received by farmers is expected to increase stock for these same items 0.2 percent in one or two years, 0.5 percent in four years and more than 2 percent in the long run" (pp. 327–328). The trend variable was robust, and the equity/liability ratio had the right sign and was significant. This can be interpreted as a sign of cyclical behavior, with higher investment in good times.

As in many empirical applications, their equations contain fewer variables than what is called for by the theory. Presumably, the equation should include all prices and a measure of technology. In general, with a short time series the empirical equation does not sustain all the pertinent variables. For instance, in the study of Heady and Tweeten (1963), the inclusion of more prices was not supported by the data. One way to deal with this problem is to collapse the prices and other exogenous variables into one measure, the rate of return. The higher the expected rate of return, the higher the investment demand. The rate of return can be thought of as a proxy for the gap between the expected marginal productivity of capital and the rental rate, labeled as z in Equation (42).

Mundlak (1964a, Chapter 6) used a panel of farm micro data to study investment in farm structures using the accelerator formulation and demonstrated the importance of aggregating the data over time in order to eliminate the noise that exists in annual micro data. This finding is consistent with lumpy investment and is not supportive of the idea of a convex cost of adjustment function that results in a gradual adjustment. As indicated earlier, this may be typical for many investments in agriculture.

The application of firm theory to the estimation of the aggregate industry investment function overlooks the fact that the supply of capital goods is not perfectly elastic. One way to incorporate this element is to estimate the allocation of total investment to the various sectors. This is the approach taken by Mundlak, Cavallo, and Domenech (1989) for Argentina, and Coeymans and Mundlak (1993) for Chile. The differential sectoral

profitability is measured by the rate of return. The long-run elasticity with respect to the ratio of sectoral rates of return is roughly 1 in both countries.

6.10. Dynamic factor demand using duality

The empirical application of the static expectations model assumes that every year the firm recalculates its plans conditional on the new information on prices and technology. The model provides an interpretation of the flexible accelerator, and it facilitates a convenient way to estimate the adjustment pattern of the quasi-fixed factors. The empirical inference has substantive and analytic aspects. The first is judged by the economic meaning of the results, regardless of the method used to derive them. The second is more complex. For the theory to be applicable, the empirical results should be consistent with the underlying conditions of the model. For the duality to be of interest, the prices should appear as arguments in the derived factor demand, their coefficients should have the right sign, and the value function should be convex in the prices. That is, in terms of Equation (49), $A_{\tilde{q}\tilde{q}}$ should be positive definite. In what follows, we summarize findings, pertinent to our discussion, of some leading studies dealing mostly with agriculture. Some of these studies use micro data, while the others use macro data.

There is a similarity in the basic assumptions underlying the static and dynamic dual analysis. Most important is the assumption, often made regardless of the level of aggregation of the data, that the factor supply and the product demand are all perfectly elastic. Other than that, the technology is generally represented by a time trend. The term "capital" is used freely to any aggregate of capital goods. Our foregoing discussion indicates that the demand elasticity for an input depends on its production elasticity or factor share. Thus, we should expect a different demand elasticity for a single item, say machinery, than for an aggregate measure.

Epstein and Denny (1983) applied the Epstein (1981) model to the US manufacturing annual data for the period 1947–1976. This application has had an influence on the studies in agriculture, and we therefore begin by reviewing here some of its pertinent sections. The technology is represented by a cost function, and hence the value function is derived by choosing the investment that minimizes the present value of the time path of the cost of production. Because it is a minimum problem, the value function should be concave in prices, which implies that the matrix analogous to $A_{\tilde{q}\tilde{q}}$ in (49) should be negative definite. In the estimation, the symmetry in price response was imposed, but the nonnegativity condition is violated. The authors argue that the violation is statistically only marginal. Following this line of thinking, we should note that the origin is also included in the joint confidence region for the price coefficients, which means that the null hypothesis of no price response cannot be rejected. The authors are aware of this problem, but do not accept the outcome because it is inconsistent with the concept of duality underlying the analysis. This raises the question of what do we learn from superimposing a model which is rejected by the data. The cost of this procedure is that we avoid the search for the reasons of the violation of confirming duality with the given sample.

The results show that labor and capital turn out as quasi-fixed. The rate of adjustment is fast for labor, an adjustment coefficient of 0.9, which implies a closure of the gap in a little over a year. On the other hand, the rate for capital is slow, an adjustment coefficient of 0.12, which means that it takes about 8 years to close the gap. The adjustment matrix is not diagonal, implying an interaction in the adjustment of the two factors toward their steady exogenous values. The authors are disturbed by the direction of the interaction. "It implies that a 'deficient' stock of labor reduces the demand for capital" [Epstein and Denny (1983, p. 660)]. This finding is acceptable, however, with the choice of technique approach.

The own price elasticities for capital and labor are negative but small, both in the short run and the long run. The largest numerical value is the long-run elasticity of capital, which varies between -0.25 and -0.18 for the three reported years. Because the technology is represented by a cost function, output is one of the arguments of the factor demand and, as in the studies based on the primal approach, it has a much stronger influence on demand. "With respect to output changes, a different pattern emerges. The short-run labor elasticity is roughly 0.6 and the long-run is roughly 60 percent higher. Most of the changes in labor occur in the short-run. For capital, the short-run response is negligible while the long-run response is large, an output elasticity approximately equal to 1.4" [Epstein and Denny (1983, p. 662)]. This implies that in the long run labor expands at about the same rate as output but capital grows at a faster rate, which is consistent with capital deepening and also with the hypothesis that capital is a carrier of new techniques.

The authors are aware of the fact that the theory is a micro theory, but it is applied to aggregate data. There would be no difference between the micro and macro models if the firms were similar in some sense, and the micro unit would be representative of the firms in the industry. However, the conditions for this, as developed by the authors and which are similar in nature to those of linear aggregation, are stringent. In the case of the cost function, the value function should be linear and additive in k and y. Specifically, this implies no interaction between size of the firm and factor intensity, which is unlikely in the case of heterogeneous technology. The authors estimate the model under these conditions and find that "... the resulting structure failed to satisfy the regularity conditions" [Epstein and Denny (1983, p. 662)]. In passing, it should be indicated that even if the stringent conditions for aggregation were maintained, there would still be the problem of upward-slopping factor supply that would differentiate between the micro and macro studies.

Turning to agriculture, we begin with macro studies of the US agricultural sector or industries thereof. One of the earliest applications of the duality model is the study by Vasavada and Chambers (1986) of the factor demand of US agriculture.[48] The model deals with four input categories: land, labor, machinery, and materials. The results indicate that land, labor, and capital services are quasi-fixed factors, and materials are

[48] Lopez (1985b) used the cost of adjustment in studying the dynamics of the Canadian food processing industry.

variable factors. The univariate flexible accelerator hypothesis is rejected; thus the adjustment process of the various factors is interdependent. The results show a long adjustment period for capital (10 years) and labor (9 years), and a short period for land (2 years). This pattern is puzzling, but before going deep into the rationalization of the results, it is noted that the coefficients of the adjustment matrix are mostly nonsignificant. This suggests that the null hypothesis of no adjustment might not be rejected, in which case there is no response to changes in the desired values. Obviously, this is inconsistent with the fact that inputs change every year.

An inspection of the price coefficients indicates that with the exception of materials, the own-price coefficients are not significantly different from zero. Furthermore, "Because all the diagonal elements are not positive, convexity of the value function cannot be accepted" [Vasavada and Chambers (1986, p. 955)].[49]

Luh and Stefanou (1991) estimate factor demand for US agriculture in 1950–1982. Like Vasavada and Chambers (1986), they also obtain a slow convergence to long-run equilibrium values: 0.15 of the gap for capital and 0.11 for labor. Interestingly, unlike Vasavada and Chambers (1986), they find independent convergence of labor and capital. This is consistent with the idea that the equations are strongly influenced by the factor supply.

Taylor and Monson (1985) study the factor demand in the US southeastern states in the period 1949–1981. The quasi-fixed factors are land and farm machinery, which the authors refer to as capital. The variable factors are labor and materials. "Fifteen of the estimated 26 parameters are at least two times their corresponding asymptotic standard errors" (p. 5). The price coefficients have the correct signs, hence monotonicity is maintained. Convexity is largely maintained. It seems though that most of the insignificant coefficients are those of prices, and this weakens the finding on convexity. The price elasticities, both short-run and long-run, are mostly low and fairly distant from 1. The hypotheses of independent rates of adjustment and instantaneous adjustment are rejected. The rate of adjustment was 0.55 for machinery and 0.18 for land, which means that it takes roughly two years to close the gap in machinery and six years to close the gap in land.

Howard and Shumway (1988) study the US dairy industry in the period 1951–1982. The analysis deals with two quasi-fixed inputs: herd size and labor, whereas feeds is a variable input. They use a modified version of the generalized Leontief equation. Their untested justification for the use of a micro model to the study of the industry is basi-

[49] Vasavada and Chambers (1986) remark that "... [t]here are no estimated diagonal elements with negative point estimates whose asymptotic confidence intervals do not encompass zero and positive numbers at traditionally reasonable levels of significance. Hence, the divergence from convexity, if it exists, may not be significant" (p. 955). This is not a strong supporting argument. It can be conjectured that if a joint confidence region were constructed for all the diagonal parameters in question, it would contain the origin, implying that the quadratic term in prices can be omitted; this reduces the model to absurdity.

cally the assumption that the technology is invariant to the size of the firm.[50] A similar assumption was tested by Epstein and Denny (1983) and was rejected.

As to the results, "Nearly half of the parameters were significant at the 5 percent level, which was quite robust compared to other estimated dynamic dual models" [Howard and Shumway (1988, p. 842)]. This is hardly a complimentary comment, and it illustrates the difficulties associated with the application of the model. R^2 is high for the inputs but low for the output (0.29).[51] The adjustment rate for cows and labor is 0.09 and 0.4 respectively. This raises a question: When prices change, why would labor respond when the adjustment in herd size is sluggish? It is suggested that "[t]he slow adjustment of cows is consistent with the very inelastic short-run milk supply found in previous studies" [Howard and Shumway (1988, p. 842)]. This now suggests that the capital stock is a function of output, but this is an explanation that the present model intends to replace and as such it is questionable. A different line of reasoning suggests that because the study deals with the industry as a whole, the changes in output reflect expected changes in aggregate demand, and this possibility is not accommodated by the model.

The monotonicity conditions on the value function were held at nearly all observations. However when the convexity was imposed, the model did not converge, and this is an indication of inconsistencies. "All the short-run own price input demand elasticities were negative, but the output own-price elasticity was positive for only fifteen of the thirty-two observations" [Howard and Shumway (1988, p. 844)]. In dynamic models, a sign reversal can happen in the short run, but this would have to come from a sign reversal in some inputs. This is not shown to be the case here. "The short-run, own-price input demand elasticities for cows and labor became more elastic over time. The increasing own-price elasticity for labor was consistent with the increasing proportion of hired to family labor over the period" (Ibid., p. 845). Again, the question of identification comes up. With what we know about the declining number of farm operators in the US (as elsewhere), the question is whether this is not a reflection of changes in labor supply rather than in labor demand.

Next, we review two studies that extend the assumption of the model to allow for a difference in the pace of adjustment between positive and negative investment. Chang and Stefanou (1988) apply the model to a panel data of 173 Pennsylvania dairy farms in 1982–1984. Hired labor and feeds are variable inputs, whereas family labor, herd size (cow), real estate, and equipment are quasi-fixed. Results are reported only for the adjustment coefficients, so that we cannot evaluate the impact of the specification on

[50] "The dairy industry consists of many price-taking firms, and theory suggests that in long run competitive equilibrium all such firms operate at the minimum average cost ... it is necessary and sufficient for consistent aggregation across firms that the value function be affine in capital" [Howard and Shumway (1988, p. 840)].

[51] The actual empirical equation is not presented. However, in general, the inputs are regressed on their lagged values and the other variables. When the dependent variable is quasi-fixed, the regression is of a stock variable on its lagged value. Such equations in general show a very good fit. The output is a flow variable, and this may explain its relatively low value.

the price coefficients. It is stated that "... at least half of the parameter estimates are significant at the 10 percent significant level especially those associated with prices of *variable* factors" (p. 149, italics by YM). If this statement suggests low precision of the estimated price coefficients, as in the other studies, the results are better for the adjustment coefficients, where most of the own adjustment coefficients are significant at the 1 percent level. The adjustment of the four quasi-fixed inputs are interdependent. There is a difference in the response when asymmetry is allowed for. "In the symmetric model, the estimated own adjustment coefficient for durable equipment is 0.8072, the highest among four quasi-fixed factors. The adjustment rates for family labor, herd size and real estate are relatively more sluggish. In the asymmetric specification the adjustment of durable equipment also appears to be sluggish. Family labor and herd size follow a similar adjustment pattern in that the contracting adjustment rate is higher than the expanding one... The adjustment rates for real estate and durable equipment are somewhat confusing in terms of their signs and magnitude" (p. 151).

Lansink and Stefanou (1997) extend further the asymmetric model by allowing also for changes in the investment regime. The model is applied to a sample of specialized cash crop farms in Holland, 1971–1992. There are 4,040 observations, 2.4 percent of which reported negative investment, 29.4 percent of which had zero investment, and the remainder of which had positive investment. Quasi-fixed inputs are machinery and rootcrop-specific area. Fixed inputs are the total area of rootcrops and other outputs and labor. There are two outputs, rootcrops and 'others'. Variable inputs include pesticides, fertilizers, and 'others'.

"This model contains 92 parameters, including two parameters related to the expected error terms in Equation (20). The estimated model generated 49 percent of the parameters estimated significant at the critical 5 percent level. Convexity ... is found not to hold" [Lansink and Stefanou (1997, p. 1346)]. It is concluded that the parameter difference between the two regimes is significant for the adjustment parameter of machinery and the parameter relating machinery investment to the quality of labor. Simulation shows response to prices in both the probability of being in a particular regime and in the magnitude.

Finally, "The adjustment rate for machinery is 13 percent a year toward the long-run equilibrium machinery target in the presence of a disinvestment regime and 7 percent a year in the presence of an investment regime" (p. 1349). The rate of disinvestment is in line with conventional rates of depreciation used for machinery, which suggests disinvestment by attrition.

Under the assumption of static expectations, firms recalculate the optimal plan every year conditional on the prevailing prices and technology. But prices are subject to variations and the firms know it, so they must exercise some judgment as to the permanence of a given price regime. This brings up the question of expectations. Luh and Stefanou (1996) replace the assumption of static expectations with "nonstatic expectations", which are introduced by first order autoregressive regressions. The model is applied to US agriculture, using two alternative data sets. The quasi-fixed inputs are capital and labor. The results are not invariant to the data set. The hypotheses of static

price expectations are all soundly rejected for one data set but not for the other. Similarly, the test for independent adjustment rejects the null (independence) for one set but not the other. Quasi-fixity is accepted for both sets. As to the rate of adjustment: "While estimated adjustment rates vary, taken together these results suggest that capital and labor take two to three years to adjust to their long-run equilibrium levels. Other adjustment cost models for US agriculture ... report adjustment rates for capital and labor ranging, respectively, from 9 percent to 55 percent and from 7 percent to 40 percent. Our study predicts moderate adjustment speed for capital but much faster labor adjustment compared to other studies" [Luh and Stefanou (1996, pp. 1001–1002)]. Not all the required properties of the value function are met (Table 6). The authors are disturbed by the fact that the results are sensitive to the data sets.

Thijssen (1996) compares static expectations with rational expectations, using panel data of Dutch dairy farms, 1970–1982. The specification is different from the studies reviewed above in that labor and land are treated as exogenous; capital is the only endogenous variable. The results obtained by imposing the constraints of the rational expectations do not make sense and are inconsistent with the theory. The results with static expectations give elasticities of long-run demand for capital of 0.59, −0.45, and −0.13 for the prices of output, capital services, and variable inputs, respectively. However, the coefficients of labor and land are insignificantly different from zero.

The impact of the resource constraint on the demand of the factors that are allowed to vary can be evaluated by comparing the short-run and long-run price elasticities. Output control as a component of agricultural policy introduces another constraint. Fulginiti and Perrin (1993) and Moschini (1988) showed that production quotas on a product reduce the supply elasticities of the nonmanaged products. This can be attributed to the reduction in the scope for substitution. Richards and Jeffrey (1997) use the dynamic duality framework and data for Alberta dairy farms over the period 1975–1991. They attribute the impact to the investment that is tied up in the purchase of production quotas, which may amount to "... half of the total cost of establishing a dairy farm, may cause farmers to face a real capital constraint" (p. 555).

As to the results, monotonicity and symmetry are not rejected, but "... imposing convexity on the full four quasi-fixed inputs model caused the estimation procedure to fail to converge" (Ibid., p. 561). The model was reduced to contain only two quasi-fixed inputs, but "[a]s with the full model, the reduced model does not converge with convexity imposed parametrically. Given these results, further estimation proceeds with two quasi-fixed inputs, dairy cattle and quota licenses, with only symmetry imposed" (Ibid., p. 561). The estimated adjustment coefficients were 0.0995 for quota and 0.1556 for cattle. Obviously, the adjustment of the quotas to their long-run equilibrium is slow, and the question is whether this reflects only the demand side or, as with the studies based on industry data, the slow adjustment reflects the changes in the supply of quotas.

6.11. Discussion

We can now repeat the questions asked in our summary discussion of the static dual approach to the estimation of the production functions. These should be answered at

two levels: methodological and substantive. On the methodological level, the answer is simple: The approach provides an efficient and powerful way to discuss and formulate dynamic factor demand. Similar to the static duality framework, this assertion is true regardless of the outcome of the empirical analysis. In this respect, the claim made by some of the authors that the empirical analysis tests the validity of the competitive conditions is not accurate. The most that can be claimed for the empirical analysis is that the conducted tests are of the particular specification. A rejection of a particular specification is not a rejection of the competitive conditions.

The substantive message is more complex. Like in the static case, the essence of the duality framework is the ability to identify the technology by means of prices. It is therefore only natural that we concentrate our attention on the role of prices. The results with the dual dynamic framework are similar, if not more pronounced, to those obtained in the static case in that the convexity in prices of the value function is generally violated. Moreover, the price effect is relatively weak, and the long-run price elasticities and, of course, the short-run elasticities of the factor demands are relatively low. In some cases the whole price matrix is not significantly different from zero. All this suggests that the raison d'être of the duality model is put to question. We return to possible explanations below.

The dynamic dual approach concentrates on the behavioral equations and grossly neglects the inference on the production function itself. This is a good example of the principle of comparative advantage. The dynamic behavior indicates a gradual adjustment to the prevailing, and ever-changing, gaps between the desired long-run values of the quasi-fixed factors and their current values. This result is obtained by the inclusion of lagged values of the dependent stock variable in the empirical equation, as has been the case with the exogenous dynamics. The difference between the two approaches is that the dual dynamic model connects the adjustment coefficients to those of the value function. This can be considered the strength of the approach, but at the same time it also represents its weakness. In essence, this approach attributes the whole dynamics to the internal cost of adjustment. The empirical results show that in most cases the adjustment is sluggish, and in this respect it is also not different from those obtained under (the presumably naive approach of) exogenous dynamics.

There are many investment studies in nonagriculture with cost of adjustment. Often the empirical equation includes output as a variable. In the exogenous dynamics case, output is introduced to the model through the explicit expression of the marginal productivity of capital, and as such, the output coefficient is related to the production function, or through the cost function when the technology is represented by the cost function. On the other hand, in the endogenous dynamics models it is introduced also, and sometimes solely, through the expression for the adjustment cost, and as such it describes a completely different process than that implied from the first case. In summarizing the empirical record in nonagriculture, Chirinko (1993) notes that output performs well in explaining investment and that the performance of prices is rather weak. He also notes a lack of robustness of the results.

Can all these results be rationalized? There are two aspects of the decision to invest in any given year: growth and timing. The growth aspect reflects the long-term view about the prospects of the contemplated investment. The question is when to act. The timing aspect is related to the prevailing price variability which generates opportunities for cost reduction, or capital gains. This possibility is ruled out in a world of static expectations where the current prices are assumed to remain constant indefinitely. This is the reason that the value function can be formulated in terms of the annual capital charges (rental rates) rather than in terms of the total expenditures on the capital goods. It is only under the latter formulation that the expected capital gains constitute a component of the rental rate, as for instance in (32) or (34). The prospects for capital gains introduce cyclical considerations into the model. This also holds true for the interest rate which varies over time and also across individuals, reflecting their financial position. However, the interest rate is taken to be constant, as in the empirical studies reviewed above.

Furthermore there is the problem of price expectations. There are no clear-cut systematic differences in the estimates associated with different assumptions about the nature of the price expectations. It is difficult to conceive that the expectations do not matter, so it must follow that the tried alternatives have something in common, probably an error component. When the price variables are subject to measurement error, their estimated coefficients are likely to be biased downward. This problem is more serious for the capital goods than for the variable inputs because they require price forecasts for the entire lifetime of the project. If this argument is true, the own price elasticities of the variable inputs should have a lower downward bias and also be more precise (have higher t-ratios) than those of the durable inputs. A superficial inspection of the studies reviewed above indicates that this might be the case.

Duality is a micro theory, and therefore the applications with macro data add additional problems. The question of whether the macro function can be considered as that of the representative firm has already been mentioned above. But the test of the conditions for the ideal aggregation that will allow this interpretation deals only with the consequences of aggregation. There is still the problem that the factor supply and product demand are not perfectly elastic as the model assumes. Consequently, there is an identification problem, and the estimated coefficients reflect both supply and demand. This problem is shared also with the static estimates, but the dynamic model has an additional problem in that the pace of the closure of the gap is likely to reflect the pace of the changes in the factor supply or product demand. For instance, in interpreting the studies on US agriculture it is important to note that the movement of labor and capital have taken opposite directions. The decision of labor to leave agriculture is a decision made by households on their employment conditional on the opportunities outside agriculture. As for capital, its supply is not perfectly elastic, and agriculture has to compete with other industries for resources. This is consistent with the study by Lee and Chambers (1986), which tests for the credit constraint in US agriculture in 1947–1980 and concludes that farmers do not face a perfectly elastic supply of funds or credit (p. 865). As such, it is also supportive of the discussion on the choice of technique.

For the micro data, we noted that in many cases the investments are lumpy. A tractor is not purchased gradually, a piece at a time, and similarly for a milking shed. This pattern is masked in the analysis with macro data because the aggregation over firms gives a smooth time path of the investment, but the results do not shed light on the decisions made at the farm level.

Why does output perform better in studies where it appears as a variable? The foregoing discussion suggested some reasons for a revealed weak price response. In addition, as illustrated in the foregoing discussion, a change in price has direct and indirect effects on the desired capital stock, and the indirect effect is considerably larger. Thus, part of the effect of output on the desired capital stock may reflect an indirect effect of price. In addition, changes in output represent not only price effects but also changes in technology. As technology is the engine of growth, it probably plays a key role in explaining actual investment in many cases.

In conclusion, the endogenous dynamics models have two basic limitations. First, they describe a dynamic process in terms of unobserved variables, and thereby lose the main potential of explaining the timing of investment; and second, their only engine for the dynamics is the internal cost of adjustment. There has been no obvious advantage to their performance in empirical analysis nor has there been any particular insight gained by their empirical application.

7. The scope for policy evalnation

In the discussion of duality, the question was raised as to where we go from here. At this stage, it is clear that this question should be addressed within the broader framework that has evolved from the foregoing discussion. The core of the production structure, as outlined above, can be summarized by the following functions:

$y(v, k, T)$	Production function	(54)
$v(p, w, k, T)$	Demand for nondurable inputs	(55)
$w\big(v, s(v)\big)$	Supply of nondurable inputs	(56)
$k^*\big(s(k^*)\big)$	Capital demand on the optimal path	(57)
$k\big(k^*, s(k)\big)$	Actual capital	(58)
$T\big(s(T)\big)$	Implemented technology	(59)

where $s(x)$ is the vector of the exogenous variables pertinent to the supply or demand of x, whichever the case may be. Specifically, $s(v)$ are the exogenous variables that affect the supply of the nondurable (variable) inputs, $s(k^*)$ affect the capital demand on the optimal path, $s(k)$ are the variables that determine the dynamics of convergence of the capital stock to the optimal path, and $s(T)$ determine the implemented technology. Some of the exogenous variables were discussed explicitly above, others are discussed in the references or are left in an implicit form. In passing we note that the role of these

variables in empirical analysis is still to be more fully unveiled in future research. The system should also include land which is not dealt with explicitly here because we have already covered considerable ground without land. Mechanically, we can think of land as being a component of capital, in which case the supply condition of this component should be carefully specified.[52]

To obtain the dynamics of the supply, substitute the functions (55)–(59) in the production function to obtain

$$y\big[p, w, s(v), s(k^*), s(k), s(T)\big]. \tag{60}$$

Obviously, a function of the form $y(p, w)$ cannot capture all the complexities of Equation (60). The function serves as an approximation whose quality depends on the importance of the missing exogenous variables, which in turn depend on the data base. More generally, this is the problem of estimates based on duality which depend heavily on prices. When dealing with micro data with constant technology, the only relevant issue that will differentiate between the general expression in (60) and $y(p, w)$ is the handling of capital. On the other hand, when dealing with aggregate time-series data, all the exogenous variables may have an important impact.

Can such systems be evaluated empirically? The answer is positive, as has been demonstrated by Cavallo and Mundlak (1982) and Mundlak, Cavallo, and Domenech (1989) for Argentina; Coeymans and Mundlak (1993) for Chile; Lachaal and Womack (1998) for Canada; and at a lower level of aggregation, McGuirk and Mundlak (1991) for the Punjab agriculture under the Green Revolution. These studies show clearly that agriculture responds to prices following endogenous dynamics, of a different form from those discussed above, and that it takes time for the response to reach its full course.

Studying the production structure in all its complexities is both research-intensive and promising. What is the alternative? I will leave it for the reader to formulate his or her own answer. However, in thinking of an answer, we have to keep in mind that more than 70 years have passed since the work of Douglas. During this period, considerable work and ingenuity has been directed to improve the specification and the estimation method, but as we have indicated, there is no simple, robust way to describe reality. In part, the reality has many faces, and in part the researchers have many faces. As in *Rashamon*, we vary in our reports of the same phenomenon.

With this background, we can now address the cardinal question of what effect policy can have on production. Traditionally, the evaluation of the consequences of policy is limited to the examination of resource allocation. The present framework introduces an additional dimension, the determination of the implemented technology. The dependence of the implemented technology on the environment is the key factor to understanding why less-developed countries lag persistently behind the performance of developed countries. The economic environment is affected by policies, sector-specific

[52] For a discussion of land, see [Mundlak (1997)].

as well as sector-neutral. The response to changes in the economic environment is not immediate, and it is therefore important to spell out the role of the dynamics of response through resource allocation and the choice of the implemented technology. This is what the above structure does.

7.1. Summary and conclusions

We have reviewed the more important issues concerning empirical production and supply analysis with emphasis on agriculture. In order to confront aspiration with reality, we have deliberately substantiated the main arguments with explicit, and in some cases detailed, references to the reviewed studies.

The literature, spread over 50 years of research, has evolved from analysis of specific issues concerning the production function *per se* to analysis which binds together competitive conditions with the technology. Initially, the incorporation of competitive conditions dealt with static (one period) analysis, and this was extended later on to dynamic analysis. The lack of robust, and often of meaningful, results triggered a search in several directions: better precision in the estimation, an appropriate parametric form of the production function, or avoidance of a parametric presentation altogether, and ultimately the consequences of heterogeneous technology. To some extent, the different approaches have been associated with different questions asked and consequently resulted in different results, which are not always comparable. This complicates the assessment, and consequently the evaluation of a given approach is done by comparing the results with the underlying assumptions and expectations, as well as with the substantive message. This state of affairs is unsatisfying because the essence of duality is that knowing the production function, one can derive the behavioral equations and conversely, but the analysis is seldom carried out that far. Still, the search in the various directions has been essential for the understanding of the process, for marking the boundaries of the empirical analysis, and for developing alternative approaches that might overcome some of the difficulties. This is research.

The primal approach consisted initially of the estimation of a Cobb–Douglas production function using both micro and macro data. The main yield of these studies consists of production elasticities, a check of the prevalence of profit maximization, and a measure of economies of scale. The results have not been robust and have varied with the samples. We have provided some numerical results for the production elasticities which, on the whole, show that labor elasticity in agriculture is smaller than in nonagriculture, indicating that agriculture is more susceptible to changes in the cost of capital and less to changes in the wage rates than nonagriculture. Economies of scale have been detected mainly in strictly cross-sectional studies and are attributed to statistical bias due to the correlation between the unobserved idiosyncratic productivity and the input level, or simply the endogeneity of inputs. The main approaches to overcome this statistical bias have been the use of covariance analysis in panel data and the use of prices as instrumental variables (and more recently a combination of the two). The covariance analysis also provides a measure of the managerial ability – the idiosyncratic productivity – of the various firms (or other observation units such as a country or a region

as well as time). This measure is based on the same concept as that of the residual, or the TFP.

The extension of the analysis to production functions with richer parametric presentation offered greater flexibility in fitting the function to data, and bred expectations for more robust results. How do such extensions modify the conclusions drawn from the Cobb–Douglas model with respect to elasticities, profit maximization, and scale economies? In most cases, no comparison of the production elasticities obtained by the different functions is reported. Perhaps additional work would be required, perhaps this question has not come up, but there is also a more profound reason. The more general functions are either nonlinear in the parameters (such as the CES, or some of the quadratic functions) or contain too many parameters which leads to multicollinearity, and therefore are not easy to estimate directly. The situation is simplified considerably when the parameters are estimated from the first order conditions for profit maximization, rather than from the production function itself. This requires the imposition of profit maximization on the model. In many cases, the dependent variables are the factor shares, or a monotone function thereof. This procedure precludes the testing of the profit maximization and of economies of scale. The explanatory variables in these equations are the inputs. Thus, the essence of such extensions is to attribute the differences in the factor shares across observations to the variations in the input ratios, whereas in the Cobb–Douglas case the elasticities are constant. In many cases the variability in the input ratios in the sample is not sufficiently large to induce the observed spread in the factor shares.

Because the parametric enrichment of the specification of the production function generated the need to use the first order conditions for profit maximization in empirical analysis, it thereby eliminated the possibility of testing this hypothesis empirically. This state of affairs generated a potential scope for the nonparametric methods which offers a simple test for profit maximization. One can think of a two-stage analysis: a preliminary test of the hypothesis by nonparametric methods, and if the hypothesis is not rejected, a follow-up with parametric specification that imposes the conditions for profit maximization. Unfortunately, this course of action suffers from the fact that under technical change, the test for profit maximization loses much of its purity. The allowance for technical change implicitly utilizes profit maximization, and thus the analysis loses not only its purity but also much of its usefulness. Having said this, we note that there is a more profound consideration. The question of profit maximization is not a qualitative one that can be answered yes or no. Even if profit maximization is the rule, there are deviations from the first order conditions, and therefore the imposition of these conditions in the estimation may lead to erroneous results. Such deviations from the first order conditions may reflect considerations such as risk, dynamic considerations in the case of the price of durables, or simply a discrepancy in the price perception between the econometrician and the firms.

Given the estimates of the primal function, it is possible to calculate the elasticities of the behavioral functions, product supply and factor demand, and the value of the objective functions, profit, cost, or revenue as the case may be. Duality offers a reverse

course of action where the point of departure is the objective function. When the objective function is known, it can be used to derive the production function. In principle, there are several reasons to use duality in empirical analysis. First, it is a powerful theoretical concept. Second, prices are thought to be exogenous and therefore can be used to identify the technology, thereby overcoming the endogeneity of the inputs in the direct estimation of the production function. Third, it may provide a useful presentation of the technology. The first point is valid, but the problem is in its empirical implementation. The second point is valid only for micro data, but even then the method does not utilize all the information available for the identification of the production function, and as such it is not efficient compared to the primal estimates. The third point is valid only when the implemented technology is independent of the prices.

When the objective function is rich in parameters, the dual specification is reduced for empirical analysis by the use of the envelope theorem to yield empirical equations where the dependent variables are inputs, outputs, or factor shares. Those are regressed on the pertinent prices, time trend, and sometimes output. When the change in the use of inputs is decomposed to price, trend (a proxy for technology), and output effects, it is found that trend and output capture most of the change, whereas the role of prices is the least important. Thus the contribution of prices to the explanation of inputs or output variations is rather limited. Duality between technology and prices holds under well-defined conditions that can be tested empirically. In most studies these underlying conditions are not fully met; in particular the concavity of the cost function or the convexity of the profit function is violated. Therefore, the estimated technology is inconsistent with the basic premises of the model.

The price elasticities of factor demand and product supply are usually obtained under the assumption that producers are price takers in the product and factor markets. On the whole, the own-price elasticities are less than one. There is no uniformity in the signs of the cross elasticities, but in general, most inputs appear to be substitutes. The magnitude of the own and cross elasticities reflects in part the fact that in reality factor supplies are not perfectly elastic as the models assume, and therefore the results need not represent demand-driven substitution as is thought. This is the case with respect to elasticities related to labor, land, and capital. We further elaborate on this subject below.

With respect to other findings, interestingly, on the whole the studies based on duality do not show increasing returns to scale. Technical change, obtained by including a time trend in regressions of factor shares, is largely labor-saving, capital-using, and fertilizer-using, with the results for land being somewhat ambiguous.

The interest of agricultural economists in the behavioral functions had long preceded the work on production functions. The work on supply response, which was triggered by policy considerations rather than methodological innovations, is similar in nature to that of the empirical estimation of behavioral functions that emerged from the estimation of the dual functions. The initial work on supply response was in part intuitive; it lacked the duality framework, and basically it had been inspired by the primal approach. Still, it emphasized two related cardinal topics whose importance has not diminished: quasi-fixed factors and dynamics. The root of the importance of these topics is in the

fact that static analysis is timeless, whereas data are dated. This requires that behavioral equations will be conditional on the available quantities of quasi-fixed factors, a condition that has been overlooked in many (but not all) of the studies based on duality. Such functions are termed short-run, or restricted, functions. Supply elasticities derived from short-run functions are inversely related to the relative importance of quasi-fixed factors (as measured by their factor shares). The larger is the relative weight of the quasi-fixed factors, the larger is the gap between the short- and long-run supply (or factor demand) elasticities. This gap was well highlighted by distributed lag analysis which introduced dynamics into the empirical analysis. The distributed lags model is a powerful empirical tool because of its simplicity. But when the distributed lags model is applied to the outputs or inputs that are endogenous in the short run, this simplicity is achieved at the cost of ignoring the underlying production structure. The extension of the analysis to the long run requires determining the optimal level of the quasi-fixed factors, and this is done within the framework of multiperiod optimization, an important subject of current research.

Intertemporal optimization determines the optimal time path for durable goods, or simply capital goods. The first-order conditions for optimization using the primal approach sets the marginal productivity of capital equal to the user cost at any point on the optimal path. Endogenous dynamics are generated within the model, mostly by the inclusion of adjustment costs, whereas exogenous dynamics superimpose the dynamics on the model without an explicit expression for the causality. Under the dual approach, as in the static case, the value function is specified parametrically and serves as a starting point for deriving the factor demand. There is a similarity in the basic appearance of the empirical equations of these alternative approaches in that they all express the capital demand in terms of incentives and the existing capital stock.

There are two aspects of the decision to invest in any given year: growth and timing. The growth aspect reflects the long-term view about the prospects of the contemplated investment, and the timing aspect is related to the question of when to act. The expected profitability of investment is affected by changes in technology and prices. Over the long haul, technology changes more than real prices. In fact, in the case of agriculture, investment has taken place in spite of a decline in real prices. Yet, the emphasis in empirical analysis has been to explain investment in terms of prices, while technology is represented by time trend. This is particularly true for studies based on the dual approach. Time trend is not sufficiently reflective of the changes in technology. Thus it might be more promising to measure the incentives in terms of the rate of return on capital, which summarizes the information on technology and prices, rather than in terms of prices.

The dynamic dual approach provides an efficient and powerful way to discuss and formulate dynamic factor demand. However, the results with the dual dynamic framework are similar, if not more pronounced, to those obtained in the static case in that convexity in prices of the value function is generally violated. Moreover, the price effect is relatively weak, and the price elasticities, and especially the short-run elasticities, of the factor demands are relatively low. In some cases the whole price matrix is not signifi-

cantly different from zero. All this suggests that the *raison d'être* of the duality model is put to question. On the other hand, in studies which, for whatever reason, include output as an explanatory variable, output appears as a very prominent variable. It is suggested that this is due to the fact that output is a good proxy for profitability and may reflect the effect of technical change, as well as prices.

There are several possible reasons for the poor performance of prices. Some of them are due to the fact that duality is a micro theory, and therefore the applications with macro data add additional problems. In addition, there is the problem of long horizon which requires generating expected prices, and technology for that matter, for the entire lifetime of the investment. These have to be generated, and there is considerable scope for error. In addition, in the case of the dual approach, the specification is very parameter-intensive, and this creates imprecision in the estimation of individual coefficients.

The empirical results indicate a gradual and sluggish adjustment to the ever changing gaps between the desired long-run values of the quasi-fixed factors and their current values. This raises a question whether the sluggish response is the outcome of the internal cost of adjustment or alternatively a reflection of the fact that total resources are limited and the economy is facing an upward-slopping factor supply, which may be fairly inelastic.

As we progress with the review, it has become evident that some of the difficulties that have been encountered in the empirical work could be accounted for if we allow for heterogeneous technology. Changes in the available technology and in the economic environment generate opportunities for firms to seize on. The implementation of new available technologies is governed by economic considerations and is affected by the variables used in conventional analysis, such as prices or capital. It is suggested that the scope of this approach should be further investigated as a step in our attempt to come up with a uniform and robust framework that would be applicable to a wide range of economic and physical environments. An important advantage of this framework is that it provides a channel for introducing the direct effect of policy on productivity.

To conclude, in spite of all these difficulties of obtaining a uniform robust model, we know today quite a bit about orders of magnitude of some important parameters.

Acknowledgement

I am indebted to Rita Butzer for comments and for editorial assistance.

References

Afriat, S.N. (1972), "Efficiency estimation of production function", International Economic Review 13(3):568–598.

Antle, J.M. (1983), "Infrastructure and aggregate agricultural productivity: International evidence", Economic Development and Cultural Change 31(3):609–619.

Antle, J.M. (1984), "The structure of US agricultural technology, 1910–78", American Journal of Agricultural Economics 66(4):414–421.

Arrow, K.J., H.B. Chenery, B.S. Minhas and R.M. Solow (1961), "Capital-labor substitution and economic efficiency", Review of Economics & Statistics 43(3):225–250.

Askari, H., and J.T. Cummings (1976), Agricultural Supply Response: A Survey of the Econometric Evidence (Praeger Publishers, New York).

Ball, V.E. (1985), "Output, input and productivity measurement in US agriculture: 1948–79", American Journal of Agricultural Economics 67(3):475–486.

Ball, V.E., J.-C. Bureau, K. Elkin and A. Somwaru (1993a), "Implications of the common agricultural policy reform: An analytical approach", USDA, mimeograph.

Ball, V.E., J.-C. Bureau, J. Butault and H.P. Witzke (1993b), "The stock of capital in European community agriculture", European Review of Agricultural Economics 20:437–450.

Ball, V.E., J.-C. Bureau, R. Nehring and A. Somwaru (1997), "Agricultural productivity revisited", American Journal of Agricultural Economics 79(4):1045–1063.

Bar-Shira, Z., and I. Finkelshtain (1999), "Simple nonparametric tests of technological change: Theory and application to US agriculture", American Journal of Agricultural Economics (forthcoming).

Bean, L.H. (1929), "The farmers' response to price", Journal of Farm Economics 11:368–385.

Bhattacharjee, J.P. (1955), "Resource use and productivity in world agriculture", Journal of Farm Economics 37(1):57–71.

Binswanger, H.P. (1974), "A cost function approach to the measurement of elasticities of factor demand and elasticities of substitution", American Journal of Agricultural Economics 56(2):377–386.

Bouchet, F., D. Orden and G.W. Norton (1989), "Sources of growth in French agriculture", American Journal of Agricultural Economics 71(2):280–293.

Brandow, G.E. (1962), "Demand for factors and supply of output in a perfectly competitive industry", Journal of Farm Economics 44(3):895–899.

Brewster, J.M., and H.L. Parsons (1946), "Can prices allocate resources in American agriculture?", Journal of Farm Economics 28(4):938–960.

Bronfenbrenner, M. (1944), "Production functions: Cobb–Douglas, interfirm, intrafirm", Econometrica 12(1):35–44.

Capalbo, S.M. (1988), "A comparison of econometric models of US agricultural productivity and aggregate technology", in: S.M. Capalbo and J.M. Antle, eds., Agricultural Productivity: Measurement and Explanation (Resources for the Future, Washington, DC) 159–188.

Capalbo, S.M., and T.T. Vo (1988), "A review of the evidence on agricultural productivity and aggregate technology", in: S.M. Capalbo and J.M. Antle, eds., Agricultural Productivity: Measurement and Explanation (Resources for the Future, Washington, DC) 96–137.

Cassels, J.M. (1933), "The nature of statistical supply curves", Journal of Farm Economics 15(2):378–387.

Cavallo, D. (1976), "A note on consistent estimation of a production function with partially transmitted errors", Department of Economics mimeo (Harvard University, Cambridge, MA).

Cavallo, D., and Y. Mundlak (1982), Agriculture and Economic Growth in an Open Economy: The Case of Argentina, Research Report No. 36 (International Food Policy Research Institute, Washington, DC).

Caves, D.W., and L.R. Christensen (1980), "Global properties of flexible functional forms", American Economic Review 70(3):422–432.

Chalfant, J.A. (1984), "Comparison of alternative functional forms with application to agricultural input data", American Journal of Agricultural Economics 66(2):216–220.

Chalfant, J.A., and B. Zhang (1997), "Variations on invariance or some unpleasant nonparametric arithmetic", American Journal of Agricultural Economics 79(4):1164–1176.

Chambers, R.G., and R.E. Just (1989), "Estimating multioutput technologies", American Journal of Agricultural Economics 71(4):980–995.

Chambers, R.G., and R.D. Pope (1994), "A virtually ideal production system: specifying and estimating the VIPS model", American Journal of Agricultural Economics 76(1):105–113.

Chang, C.C., and S.E. Stefanou (1988), "Specification and estimation of asymmetric adjustment rates for quasi-fixed factors of production", Journal of Economic Dynamics and Control 12(1):145–151.

Chavas, J.-P. (1994), "Production and investment decisions under sunk cost and temporal uncertainty", American Journal of Agricultural Economics 76(1):114–127.

Chavas, J.-P., and T.L. Cox (1988), "A nonparametric analysis of agricultural technology", American Journal of Agricultural Economics 70(2):303–310.

Chavas, J.-P., and T.L. Cox (1994), "A primal-dual approach to nonparametric productivity analysis: The case of US agriculture", The Journal of Productivity Analysis 5(4):359–373.

Chavas, J.-P., and T.L. Cox (1992), "A nonparametric analysis of agricultural technology", American Journal of Agricultural Economics 74:583–591.

Chenery, H.B. (1952), "Overcapacity and the acceleration principle", Econometrica 20(1):1–28.

Chennareddy, V. (1967), "Production efficiency in South Indian agriculture", Journal of Farm Economics 49(4):816–820.

Chirinko, R.S. (1993), "Business fixed investment spending: a critical survey of modeling strategies, empirical results, and policy implications", Journal of Economic Literature 31(4):1875–1911.

Christensen, L.R., D.W. Jorgenson and L.J. Lau (1973), "Transcendental logarithmic production frontiers", Review of Economics and Statistics 55(1):28–45.

Clark, C. (1973), The Value of Agricultural Land (Pergamon Press, Oxford).

Clark, J.S., and C.E. Youngblood (1992), "Estimating duality models with biased technical change: A time series approach", American Journal of Agricultural Economics 74(2):353–360.

Cobb, C.W., and P.H. Douglas (1928), "A theory of production", American Economic Review 18(1):139–165.

Coeymans, J.E., and Y. Mundlak (1993), Sectoral Growth in Chile: 1962–82, Research Report No. 95 (International Food Policy Research Institute, Washington, DC).

Colyer, D., and G. Jimenez (1971), "Supervised credit as a tool in agricultural development", American Journal of Agricultural Economics 53(4):639–642.

Cox, T.L., and J.-P. Chavas (1990), "A nonparametric analysis of productivity: The case of US agriculture", European Review of Agricultural Economics 17(4):449–464.

Diewert, W.E. (1974), "Applications of duality theory", in: M. Intrilligator and D.A. Kendrick, eds., Frontiers of Quantitative Economics, Vol. II (North-Holland, Amsterdam).

Douglas, P.H. (1976), "The Cobb–Douglas production function once again: its history. Its testing, and some new empirical values", Journal of Political Economy 84(5):903–916.

Edwards, C. (1959), "Resource fixity and farm organization", Journal of Farm Economics 41(4):747–759.

Epstein, L.G. (1981), "Duality theory and functional forms for dynamic factor demands", Review of Economic Studies 48(I):81–95.

Epstein, L.G., and M.G.S. Denny (1983), "The multivariate flexible accelerator model: its empirical restrictions and an application to US manufacturing", Econometrica 51(3):647–674.

Evenson, R.E., and Y. Kislev (1975), Agricultural Research and Productivity (Yale University Press, New Haven).

Fawson, C., and C.R. Shumway (1988), "A nonparametric investigation of agricultural production behavior for US subregions", American Journal of Agricultural Economics 70(2):311–317.

Featherstone, A.M., G.A. Moghnieh and B.K. Goodwin (1995), "Farm-level nonparametric analysis of cost-minimization and profit-maximization behavior", Agricultural Economics 13(2):109–117.

Floyd, J.E. (1965), "The effects of farm price supports on the returns to land and labor in agriculture", Journal of Political Economy 73(2):148–158.

Fox, G., and L. Kivanda (1994), "Popper or production?", Canadian Journal of Agricultural Economics 42(I):1–13.

Fulginiti, L., and R. Perrin (1993), "The theory and measurement of producer response under quotas", Review of Economic and Statistics 75(1):97–106.

Fuss, M., and D. McFadden (1978), "Flexibility versus efficiency in ex ante plant design", in: M. Fuss and D. McFadden, eds., Production Economics: A Dual Approach to Theory and Applications (North-Holland, Amsterdam) 311–364.

Fuss, M., D. McFadden and Y. Mundlak (1978), "A survey of functional forms in the economic analysis of production", in: M. Fuss and D. McFadden, eds., Production Economics: A Dual Approach to Theory and Applications (North-Holland, Amsterdam) 219–268.

Galbraith, J.K., and J.D. Black (1938), "The maintenance of agricultural production during depression: The explanations reviewed", Journal of Political Economy 46(3):305–323.

Galeotti, M. (1996), "The intertemporal dimension of neoclassical production theory", Journal of Economic Surveys 10(4):421–460.

Gould, J.P. (1968), "Adjustment costs in the theory of investment of the firm", Review of Economic Studies 35(1):47–55.

Griliches, Z. (1959), "The demand for inputs in agriculture and a derived supply elasticity", Journal of Farm Economics 41(2):309–322.

Griliches, Z. (1960), "The demand for a durable input: Farm tractors in the United States, 1921–57", in: A.C. Harberger, ed., The Demand for Durable Goods (The University of Chicago Press, Chicago) 181–207.

Griliches, Z. (1963a), "The sources of measured productivity growth: United States agriculture, 1940–1960", Journal of Political Economy 71(4):331–346.

Griliches, Z. (1963b), "Estimates of the aggregate agricultural production function from cross-sectional data", Journal of Farm Economics 45(2):419–428.

Griliches, Z. (1963c), "Capital stock in investment functions: Some problems of concept and measurement", in: C.F. Christ et al., eds., Measurement in Economics: Studies in Mathematical Economics and Econometrics, in Memory of Yehuda Grunfeld (Stanford University Press, Stanford, CA).

Griliches, Z. (1964), "Research expenditures, education, and the aggregate agricultural production function", American Economic Review 54(6):961–974.

Griliches, Z., and D. Jorgenson (1966), "Sources of measured productivity change: Capital input", American Economic Review 56(2):50–61.

Haavelmo, T. (1947), "Methods of measuring the marginal propensity to consume", Journal of the American Statistical Association 42:105–122.

Hall, R.E. (1973), "The specification of technology with several kinds of output", Journal of Political Economy 81(4):878–892.

Hamermesh, D.S., and G.A. Pfann (1996), "Adjustment costs in factor demand", Journal of Economic Literature 34(3):1264–1292.

Hanoch, G. (1975), "Production and demand models with direct or indirect implicit additivity", Econometrica 43(3):395–419.

Hanoch, G., and M. Rothschild (1972), "Testing the assumptions of production theory: A nonparametric approach", Journal of Political Economics 80:256–75.

Hayami, Y. (1969), "Sources of agricultural productivity gap among selected countries", American Journal of Agricultural Economics 51(3):564–575.

Hayami, Y. (1970), "On the use of the Cobb–Douglas production function on the cross-country analysis of agricultural production", American Journal of Agricultural Economics 52(2):327–329.

Hayami, Y., and V.W. Ruttan (1970), "Agricultural productivity differences among countries", American Economic Review 60(5):895–911.

Headley, J.C. (1968), "Estimating the productivity of agricultural pesticides", American Journal of Agricultural Economics 50(1):13–23.

Heady, E.O. (1946), "Production functions from a random sample of farms", Journal of Farm Economics 28(4):989–1004.

Heady, E.O., and J.L. Dillon (1961), Agricultural Production Functions (Iowa State University Press, Ames).

Heady, E.O., and L.G. Tweeten (1963), Resource Demand and the Structure of the Agricultural Industry (Iowa State University Press, Ames).

Herdt, R.W. (1971), "Resource productivity in Indian agriculture", American Journal of Agricultural Economics 53(3):517–521.

Hildebrand, J.R. (1960), "Some difficulties with empirical results from whole-farm Cobb–Douglas-type production functions", Journal of Farm Economics 42(4):897–904.

Hoch, I. (1955), "Estimation of production function parameters and testing for efficiency", Econometrica 23(3):325–326.

Hoch, I. (1958), "Simultaneous equation bias in the context of the Cobb–Douglas production function", Econometrica 26(4):566–578.

Hoch, I. (1962), "Estimation of production function parameters combining time-series and cross-section data", Econometrica 30(1):34–53.

Hopper, W.D. (1965), "Allocation efficiency in a traditional Indian agriculture", Journal of Farm Economics 47(3):611–624.

Howard, W.H., and C.R. Shumway (1988), "Dynamic adjustment in the US dairy industry", American Journal of Agricultural Economics 70(4):837–847.

Huang, Y. (1971), "Allocation efficiency in a developing agricultural economy in Malaya", American Journal of Agricultural Economics 53(3):514–516.

Huffman, W.E., and R.E. Evenson (1989), "Supply and demand functions for multiproduct US cash grain farms: Biases caused by research and other policies", American Journal of Agricultural Economics 71(3):761–773.

Johnson, D.G. (1950), "The nature of supply function for agricultural products", American Economic Review 40(4):539–564.

Johnson, G.L. (1958), "Supply function-some facts and notions", in: E.O. Heady, H.G. Diesslin, H.R. Jensen, and G.L. Johnson, eds., Agricultural Adjustment Problems in a Growing Economy (The Iowa State College Press, Ames, Iowa) 74–93.

Johnson, G.L., and L. Quance (1972), The Overproduction Trap in US Agriculture (The Johns Hopkins University Press for Resources for the Future, Baltimore, Maryland).

Jorgenson, D.W. (1963), "Capital theory and investment behavior", American Economic Review 53(2):247–259.

Jorgenson, D.W. (1967), "The theory of investment behavior", in: R. Ferber, ed., Determinants of Investment Behavior, (Columbia University Press, National Bureau of Economic Research, New York) 129–155.

Jorgenson, D.W. (1986), "Econometric methods for modeling producer behavior", in: Z. Griliches and M.D. Intriligator, eds., Handbook of Econometrics, Vol. III (North-Holland, Amsterdam) 1841–1915.

Jorgenson, D.W., and F.M. Gollop (1992), "Productivity Growth in US agriculture: A postwar perspective", American Journal of Agricultural Economics 74(3):745–750.

Just, R.E., D. Zilberman and E. Hochman (1983), "Estimation of multicrop production functions", American Journal of Agricultural Economics 65(4):770–780.

Kako, T. (1978), "Decomposition analysis of derived demand for factor inputs: The case of rice production in Japan", American Journal of Agricultural Economics 60(4):628–635.

Kamien, M.I., and N.L. Schwartz (1991), Dynamic Optimization, 2nd edition (North-Holland, Amsterdam).

Kawagoe, T., and Y. Hayami (1983), "The production structure of world agriculture: An intercountry cross-section analysis", The Developing Economies 21:189–206.

Kawagoe, T., Y. Hayami and V.W. Ruttan (1985), "The intercountry agricultural production function and productivity differences amoug countries", Journal of Development Economics 19:113–132.

Kislev, Y. (1966), "Overestimates of returns to scale in agriculture – a case of synchronized aggregation", Journal of Farm Economics 48(4):967–983.

Kislev, Y., and W. Peterson (1996), "Economies of scale in agriculture: A reexamination of evidence", in: J.M. Antle and D. Sumner, eds., The Economics of Agriculture, Papers in honor of D. Gale Johnson, Vol. 2 (Uuiversity of Chicago Press, Chicago) 156–170.

Klein, L.R. (1953), A Textbook of Econometrics (Row, Peterson and Co, Evanston, IL).

Koyck, L. (1954), Distributed Lags and Investment Analysis (North-Holland, Amsterdam).

Kuroda, Y. (1987), "The production structure and demand for labor in postwar Japanese agriculture, 1952–82", American Journal of Agricultural Economics 69(2):328–337.

Lachaal, L., and A.W. Womack (1998), "Impacts of trade and macroeconomic linkages on Canadian agriculture", American Journal of Agricultural Economics 80(3):534–542.

Lansink, A.O., and S.E. Stefanou (1997), "Asymmetric adjustment of dynamic factors at the firm level", American Journal of Agricultural Economics 79(4):1340–1351.

Larson, D., R. Butzer, Y. Mundlak and A. Crego (1999), "A cross-country database for sector investment and capital," Working Paper No. 9903 (The Center for Agricultural Economic Research, Rehovot).

Lau, L.J. (1978), "Applications of profit functions", in: M. Fuss and D. McFadden, eds., Production Economics: A Dual Approach to Theory and Applications, Vol. 1 (North-Holland, Amsterdam) 133–216.

Lau, L.J., and P.A. Yotopoulos (1972), "Profit, supply, and factor demand functions", American Journal of Agricultural Economics 54(1):11–18.

Leathers, H.D. (1991), "Allocable fixed inputs as a cause of joiut production: A cost function approach", American Journal of Agricultural Economics 73(4):1083–1090.

Lee, H., and R.G. Chambers (1986), "Expenditure constraints and profit maximization in US agriculture", American Journal of Agricultural Economics 68(4):857–865.

Lopez, R.E. (1980), "The structure of production and the derived demand for inputs in Canadian agriculture", American Journal of Agricultural Economics 62(1):38–45.

Lopez, R.E. (1984), "Estimating substitution and expansion effects using a profit function framework", American Journal of Agricultural Economics 66(3):358–367.

Lopez, R.E. (1985a), "Structural implications of a class of flexible functional forms for profit functions", International Economic Review 26(3):593–601.

Lopez, R.E. (1985b), "Supply response and investment in the Canadian food processing industry", American Journal of Agricultural Economics 67(1):40–48.

Lucas, R.E. (1967), "Optimal investment policy and the flexible accelerator", International Economic Review 8:78–85.

Luh, Y.H., and S.E. Stefanou (1991), "Productivity growth in US agriculture under dynamic adjustment", American Journal of Agricultural Economics 73(4):1116–1125.

Luh, Y.H., and S.E. Stefanou (1996), "Estimating dynamic dual models under nonstatic expectations", American Journal of Agricultural Economics 78(4):991–1003.

Marschak, J., and W.H. Andrews, Jr. (1944), "Random simultaneous equations and the theory of production", Econometrica 12(3&4):143–205.

McFadden, D. (1978), "Cost, revenue, and profit functions", in: M. Fuss and D. McFadden, eds., Production Economics: A Dual Approach to Theory and Applications (North-Holland, Amsterdam) 3–109.

McGuirk, A., and Y. Mundlak (1991), Incentives and Constraints in the Transformation of Punjab Agriculture, Research Report No. 87 (International Food Policy Research Institute, Washington, DC).

McLaren, K.R., and R.J. Cooper (1980), "Intertemporal duality: Application to the theory of the firm", Econometrica 48(7):1755–1762.

Mittelhammer, R.C., D.L. Young, D. Tasanasanta and J.T. Donnelly (1980), "Mitigating the effects of multicollinearity using exact and stochastic restrictions: The case of an aggregate agricultural production function in Thailand", American Journal of Agricultural Economics 62:199–210.

Moschini, G. (1988), "A model of production with supply management for the Canadian agricultural sector", American Journal of Agricultural Economics 70(2):318–329.

Mundlak, Y. (1961), "Empirical production function free of management bias", Journal of Farm Economics 43(1):44–56.

Mundlak, Y. (1963a), "Estimation of production and behavioral functions from a combination of cross-section and time-series data", in: C.F. Christ et al., eds., Measurement in Economics: Studies in Mathematical Economics and Econometrics, in Memory of Yehuda Grunfeld (Stanford University Press, Stanford, CA) 138–166.

Mundlak, Y. (1963b), "Specification and estimation of multiproduct production functions", Journal of Farm Economics 45(2):433–443.

Mundlak, Y. (1964a), An Economic Analysis of Established Family Farms in Israel, 1953–1958 (The Falk Project for Economic Research in Israel, Jerusalem).

Mundlak, Y. (1964b), "Transcendental multiproduct production functions", International Economic Review 5(3):273–284.

Mundlak, Y. (1966), "On the microeconomic theory of distributed lags", Review of Economics and Statistics 48(1):51–60.

Mundlak, Y. (1967), "Long-run coefficients and distributed lag analysis: A reformulation", Econometrica 35(2):278–293.

Mundlak, Y. (1968) "Elasticities of substitution and the theory of derived demand", Review of Economic Studies 35:225–236.

Mundlak, Y. (1988), "Endogenous technology and the measurement of productivity", in: S.M. Capalbo and J.M. Antle, eds., Agricultural Productivity: Measurement and Explanation (Resources for the Future, Washington, DC) 316–331.

Mundlak, Y. (1993), "On the empirical aspects of economic growth theory", American Economic Review 83(2):415–420.

Mundlak, Y. (1996a), "Production function estimation: Reviving the primal", Econometrica 64(2):431–438.

Mundlak, Y. (1996b) "On the aggregate agricultural supply", in: J.M. Antle and D. Sumner, eds., The Economics of Agriculture, Papers in honor of D. Gale Johnson, Vol. 2 (University of Chicago Press, Chicago) 101–120.

Mundlak, Y. (1997), "Land expansion, land augmentation, and land saving", Benjamin H. Hibbard Memorial Lecture Series (Department of Agricultural and Applied Economics, University of Wisconsin, Madison, WI).

Mundlak, Y. (2000), Agriculture and economic growth; Theory and Measurement (forthcoming).

Mundlak, Y., D. Cavallo and R. Domenech (1989), Agriculture and Economic Growth in Argentina, 1913–84, Research Report No. 76 (International Food Policy Research Institute, Washington, DC).

Mundlak, Y., and R. Hellinghausen (1982), "The intercountry agricultural production function: Another view", American Journal of Agricultural Economics 64(4):664–672.

Mundlak, Y., and I. Hoch (1965), "Consequences of alternative specifications in estimation of Cobb–Douglas production functions", Econometrica 33(4):814–828.

Mundlak, Y., D. Larson and R. Butzer (1999), "Rethinking within and between regression: The case of agricultural production functions," Annales D'Economie et de Statistique 55–56:475–501.

Mundlak, Y., and A. Razin (1969), "Aggregation, index numbers and the measurement of technical change", Review of Economics and Statistics 51(2):166–175.

Mundlak, Y., and A. Razin (1971), "On multistage multiproduct production functions", American Journal of Agricultural Economics 53(3):491–499.

Mundlak, Y., and Z. Volcani (1973), "The correspondence of efficiency frontier as a generalization of the cost function", International Economic Review 14(1):223–233.

Nadiri, M.I., and S. Rosen (1969), "Interrelated factor demand functions", American Economic Review 59(4):457–471.

Nerlove, M. (1956), "Estimates of the elasticities of supply of selected agricultural commodities", Journal of Farm Economics 38(2):496–509.

Nerlove, M. (1958), "Distributed lags and estimation of long-run supply and demand elasticities: Theoretical considerations", Journal of Farm Economics 40(2):301–313.

Nerlove, M. (1963), "Returns to scale in electricity supply", in: C.F. Christ et al., eds., Measurement in Economics: Studies in Mathematical Economics and Econometrics, in Memory of Yehuda Grunfeld (Stanford University Press, Stanford, CA) 167–200.

Nguyen, D. (1979), "On agricultural productivity differences among countries", American Journal of Agricultural Economics 61(3):565–570.

Ray, S.C. (1982), "A translog cost function analysis of US agriculture, 1939–77", American Journal of Agricultural Economics 64(3):490–498.

Reder, M.W. (1943), "An alternative interpretation of the Cobb–Douglas function", Econometrica 11(3&4):259–264.

Richards, T.J., and S.R. Jeffrey (1997), "The effect of supply management on Herd size in Alberta dairy", American Journal of Agricultural Economics 79(2):555–565.

Rosegrant, M.W., and R.E. Evenson (1992), "Agricultural productivity and sources of growth in South Asia", American Journal of Agricultural Economics 74(3):757–761.

Ruttan, V.W. (1956), "The contribution of technological progress to farm output: 1950–1975", Review of Economics and Statistics 38(1):61–69.

Sadan, E. (1968), "Capital formation and growth in the Israeli cooperative farm", American Journal of Agricultural Economics 50:975–990.

Sahota, G.S. (1968), "Efficiency of resource allocation in Indian agriculture", American Journal of Agricultural Economics 50:584–605.

Schultz, T.W. (1953), Economic Organization of Agriculture (McGraw-Hill, New York).

Shih, J.T., L.J. Hushak and N. Rask (1977), "The validity of the Cobb–Douglas specification in Taiwan's developing agriculture", American Journal of Agricultural Economics 59(3):554–558.

Shumway, C.R. (1983), "Supply, demand, and technology in a multiproduct industry: Texas field crops", American Journal of Agricultural Economics 65(4):748–760.

Shumway, C.R. (1995), "Recent duality contributions in production economics", Journal of Agricultural and Resource Economics 20(1):178–194.

Shumway, C.R., and W.P. Alexander (1988), "Agricultural product supplies and input demands: Regional comparisons", American Journal of Agricultural Economics 70(1):153–161.

Shumway, C.R., R.D. Pope and E. Nash (1984), "Allocatable fixed inputs and jointness in agricultural production: Implications for modeling", American Journal of Agricultural Economics 66(1):72–78.

Shumway, C.R., R.R. Saez and P.E. Gottret (1988), "Multiproduct supply and input demand in US agriculture", American Journal of Agricultural Economics 70(2):330–337.

Shumway, C.R., H. Talpaz and B.R. Beattie (1979), "The factor share approach to production function "estimation": Actual or estimated equilibrium shares?", American Journal of Agricultural Economics 61(3):561–564.

Smith, V.E. (1945), "The statistical production function", Quarterly Journal of Economics 59(4):543–562.

Solow, R.M. (1957), "Technical change and the aggregate production function", Review of Economics and Statistics 39(3):312–320.

Tauer, L.W. (1995), "Do New York dairy farmers maximize profits or minimize costs?", American Journal of Agricultural Economics 77(2):421–429.

Taylor, T.G., and M.J. Monson (1985), "Dynamic factor demands for aggregate southeastern United States agriculture", Southern Journal of Agricultural Economics 17(2):1–9.

Thijssen, G. (1996), "Farmers' investment behavior: An empirical assessment of two specifications of expectations", American Journal of Agricultural Economics 78(1):166–174.

Tintner, G. (1944), "A note on the derivation of production functions from farm records", Econometrica 12:26–34.

Tintner, G., and O.H. Brownlee (1944), "Production functions derived from farm records", Journal of Farm Economics 26(3):566–571 (a correction in JFE Feb. 1953, 35:123).

Treadway, A.B. (1969), "On rational entrepreneurial behavior and the demand for investment", Review of Economic Studies 36(2):227–239.

Tweeten, L.G., and L. Quance (1969), "Positivistic measures of aggregate supply elasticities: Some new approaches", American Economic Review 59(2):175–183.

Ulveling, E.F., and L.B. Fletcher (1970), "A Cobb–Douglas production function with variable returns to scale", American Journal of Agricultural Economics 52(2):322–326.

Varian, H.R. (1984), "The nonparametric approach to production analysis", Econometrica 52(3):579–597.

Vasavada, U., and R.G. Chambers (1986), "Investment in US agriculture", American Journal of Agricultural Economics 68(4):950–960.

Walters, A.A. (1963), "Production and cost functions: An econometric survey", Econometrica 31(1–2):1–66.

Weaver, R.D. (1983), "Multiple input, multiple output production choices and technology in the US wheat region", American Journal of Agricultural Economics 65(1):45–56.

Wolfson, R.J. (1958), "An econometric investigation of regional differentials in American agricultural wages", Econometrica 26(2):225–257.

Working, E.J. (1927), "What do statistical demand curves show?", Quarterly Journal of Economics 41:212–235.

Wu, C.C. (1977), "Education in farm production: The case of Taiwan", American Journal of Agricultural Economics 59(4):699–709.

Yotopoulos, P.A. (1967), Allocative Efficiency in Economic Development (Center of Planning and Economic Research, Athens).

Yotopoulos, P.A., L.J. Lau and W.L. Lin (1976), "Microeconomic output supply and factor demand functions in the agriculture of the province of Taiwan", American Journal of Agricultural Economics 58(2):333–340.

Chapter 2

UNCERTAINTY, RISK AVERSION, AND RISK MANAGEMENT FOR AGRICULTURAL PRODUCERS

GIANCARLO MOSCHINI and DAVID A. HENNESSY
Department of Economics, Iowa State University, Ames, IA

Contents

Handbook of Agricultural Economics, Volume 1, Edited by B. Gardner and G. Rausser

Abstract

Uncertainty and risk are quintessential features of agricultural production. After a brief overview of the main sources of agricultural risk, we provide an exposition of expected utility theory and of the notion of risk aversion. This is followed by a basic analysis of agricultural production decisions under risk, including some comparative statics results from stylized models. Selected empirical topics are surveyed, with emphasis on risk analyses as they pertain to production decisions at the farm level. Risk management is then discussed, and a synthesis of hedging models is presented. We conclude with a detailed review of agricultural insurance, with emphasis on the moral hazard and adverse selection problems that arise in the context of crop insurance.

JEL classification: Q12

1. Introdnction

Because of the complexities of physical and economic systems, the unfolding of most processes that we care about exhibits attributes that cannot be forecast with absolute accuracy. The immediate implication of this *uncertainty* for economic agents is that many possible outcomes are usually associated with any one chosen action. Thus, decision making under uncertainty is characterized by *risk*, because typically not all possible consequences are equally desirable. Although uncertainty and risk are ubiquitous, in agriculture they constitute an essential feature of the production environment and arguably warrant a detailed analysis.

Considerable research has been devoted to exploring questions connected with the effects of uncertainty and risk in agriculture, and these efforts have paralleled related developments in the general economics literature. In this chapter we set out to review a number of these studies, especially as they relate to farm-level production decisions. To economize on our coverage of earlier work, and at the risk of not doing justice to some ground-breaking studies, we can refer to Dillon's (1971) survey as a starting point. In addition to providing an exposition of expected utility (EU) theory, which contributed to rooting subsequent studies in modern economic analysis, that survey provides an exhaustive account of previous studies of uncertainty and risk in agricultural economics. Subsequent useful compendia include Anderson, Dillon and Hardaker (1977), who consider a comprehensive set of applications of decision theory to agricultural production under uncertainty, and Newbery and Stiglitz (1981), who not only provide a thorough study of commodity price stabilization issues, but also analyze a number of problems that are relevant to the understanding of risk in agriculture.

The aforementioned contributions have been accompanied and followed by considerable research that is relevant to our pursuit. As we undertake to provide a critical survey of these studies, we are mindful of the subjective bias and unintended oversights that an exercise such as this inevitably entails, a risk heightened in our case by the encompassing nature of the topic and the sheer volume of the relevant literature. We apologize for errors of omission and commission, and we hope that our review will nonetheless prove useful to the applied researcher.

1.1. Uncertainty and risk in agriculture

Despite the fact that any taxonomy is somewhat arbitrary, it is useful to start by outlining the main sources of uncertainty and risk that are relevant from the point of view of the agricultural producer. First, there is what can be broadly defined as *production uncertainty*: in agriculture the amount and quality of output that will result from a given bundle of inputs are typically not known with certainty, i.e., the production function is stochastic. This uncertainty is due to the fact that uncontrollable elements, such as weather, play a fundamental role in agricultural production. The effects of these uncontrollable factors are heightened by the fact that time itself plays a particularly important role in agricultural production, because long production lags are dictated by the biological processes that underlie the production of crops and the growth of animals. Although

there are parallels in other production activities, it is fair to say that production uncertainty is a quintessential feature of agricultural production.

Price uncertainty is also a standard attribute of farming activities. Because of the biological production lags mentioned above, production decisions have to be made far in advance of realizing the final product, so that the market price for the output is typically not known at the time these decisions have to be made. Price uncertainty, of course, is all the more relevant because of the inherent volatility of agricultural markets. Such volatility may be due to demand fluctuations, which are particularly important when a sizable portion of output is destined for the export market. Production uncertainty as discussed earlier, however, also contributes to price uncertainty because price needs to adjust to clear the market. In this process some typical features of agricultural markets (a large number of competitive producers, relatively homogeneous output, and inelastic demand) are responsible for generating considerable price volatility, even for moderate production shocks.

Additional sources of uncertainty are relevant to farming decisions when longer-term economic problems are considered. *Technological uncertainty*, associated with the evolution of production techniques that may make quasi-fixed past investments obsolete, emerges as a marked feature of agricultural production. Clearly, the randomness of new knowledge development affects production technologies in all sectors. What makes it perhaps more relevant to agriculture, however, is the fact that technological innovations here are the product of research and development efforts carried out elsewhere (for instance, by firms supplying inputs to agriculture), such that competitive farmers are captive players in the process. *Policy uncertainty* also plays an important role in agriculture. Again, economic policies have impacts on all sectors through their effects on such things as taxes, interest rates, exchange rates, regulation, provision of public goods, and so on. Yet, because agriculture in many countries is characterized by an intricate system of government interventions, and because the need for changing these policy interventions in recent times has remained strong (witness the recent transformation of key features of the agricultural policy of the United States and the European Union, or the emerging concerns about the environmental impacts of agricultural production), this source of uncertainty creates considerable risk for agricultural investments.

1.2. Modeling issues

Two concepts of paramount importance in economic modeling are *optimization* (the rational behavior of economic agents) and *equilibrium* (the balancing of individual claims in a market setting). The application of both of these concepts raises problematic issues when uncertainty is involved. In particular, to apply the powerful apparatus of optimization to individual choices under uncertainty one needs to determine what exactly is being optimized. Although a universally satisfactory answer to this question is far from obvious, the most widely used idea is that agents exposed to uncertainty and risk maximize expected utility. This paradigm represents the culmination of a research program that dates back to Bernoulli (1738), and rests on some compelling assumptions about

individual choice. Most of the applications that we will review rely on the EU model (indeed, often some restricted version of it). Thus, in what follows we will briefly review the EU hypothesis before we proceed with a survey of applications. We should note, however, that despite its normative appeal, the EU framework has recently come under intense scrutiny because of its inability to describe some features of individual behavior under risk, and a number of generalizations of the EU model have been proposed [Machina (1987), Quiggin (1993)].

A modeling strategy that recurs in the applied literature is the distinction between uncertainty and risk attributed to Knight (1921). According to this view, risk arises when the stochastic elements of a decision problem can be characterized in terms of numerical objective probabilities, whereas uncertainty refers to decision settings with random outcomes that lack such objective probabilities. With the widespread acceptance of probabilities as subjective beliefs, Knight's distinction between risk and uncertainty is virtually meaningless and, like other authors [e.g., Hirshleifer and Riley (1992)], we will ignore it here.[1] Thus, the notions of uncertainty and risk are interchangeable in what follows, although, like Robison and Barry (1987), we tend to use the word *uncertainty* mostly to describe the environment in which economics decisions are made, and the word *risk* to characterize the economically relevant implications of uncertainty.

2. Decision making nnder nncertainty

Economic models of individual choice are necessarily rooted in the assumption of rationality on the part of decision makers. Perhaps the most common and widely understood such model is given by the neoclassical theory of consumer choice under certainty. The primitive assumption is that there is a preference ordering on commodity bundles that satisfies the consistency requirements of completeness and transitivity. These basic rationality postulates, coupled with the assumption of continuity (a hardly avoidable and basically harmless mathematical simplification), allow consumer choices to be characterized in terms of an ordinal utility function, a construct that enhances the analytical power of the assumptions. Choice under uncertainty could be characterized within this elementary setting, given minor modification of the original assumptions. For example, as in Debreu (1959), the standard preference ordering of neoclassical consumption theory could be applied to state-contingent commodity bundles. The analysis can then proceed without reference to the probability of the various states of nature. Whereas such an approach has proven useful for some problems [Arrow (1964), Hirshleifer (1966)], for a number of other cases, including applications typically of interest to agricultural economists, a more specific framework of analysis is desirable. By explicitly recognizing the mutually exclusive nature of alternative random consequences,

[1] We should note, however, that in some cases this approach is not totally satisfactory, as illustrated for example by the so-called Ellsberg paradox [Ellsberg (1961)].

one can get a powerful representation of decision making under uncertainty. This leads to the so-called EU model of decision under uncertainty, arguably the most important achievement of modern economic analysis of individual behavior. Although there exist a number of lucid expositions of this model [for a textbook treatment, see Mas-Colell et al. (1995, Chapter 6)], we present (somewhat informally) the main features of EU theory, to set the stage for the review of applications that follows.

2.1. *Preferences over lotteries and the expected utility model*

Let A represent the set of all possible actions available to decision makers, and let S represent the set of all possible states of nature. The specific action chosen by the agent and the particular state of nature that is realized (with the former choice being made prior to the resolution of uncertainty about the true state of nature) determine the outcomes (consequences) that the agent cares about. In other words, consequences are random variables as given by the function $c: S \times A \to C$, where C is the set of all possible consequences. For example, C could be the set of all possible commodity bundles as in standard consumer theory, in which case $C = \mathbb{R}^n_+$. Alternatively, as in many applications, it is monetary outcomes that are of interest to the decision makers, in which case one can put $C = \mathbb{R}$. Suppose for simplicity that the set C is finite, and that there are N possible consequences. Given an objectively known probability for each state of nature, then choosing a particular action will result in a probability distribution (a lottery, a gamble) over outcomes. Formally, one can define a lottery as a probability list $L \equiv (\ell_1, \ell_2, \ldots, \ell_N)$ such that ℓ_i is the probability (likelihood) that consequence $c_i \in C$ will arise (of course, $\ell_i \in [0, 1]$ and $\sum_i \ell_i = 1$).

In this setting, primitive preferences are represented by a preference relation $\succsim$ defined over the set of all possible lotteries $\mathcal{L}$. Assuming that this relation is rational (complete and transitive) and satisfies a specific continuity assumption, then all lotteries can be ranked by a function $V: L \to \mathbb{R}$ in the sense that, for any two lotteries L and L', we have $L \succsim L' \Leftrightarrow V(L) \geqslant V(L')$. Because the underlying assumption is that the decision maker is concerned only with the ultimate consequences, compound lotteries in this setting are always equivalent to the corresponding reduced lottery. Thus, for example, a gamble that gives lottery L with probability λ and lottery L' with probability $(1 - \lambda)$ is equivalent to a simple lottery whose probabilities are given by the mixture $\lambda L + (1 - \lambda)L''$. So far, the parallel with standard consumer theory is quite close [in particular, for example, $V(L)$ is an ordinal function]. To get the EU model, a further assumption is required at this point, namely the "independence axiom" [Samuelson (1952)]. This condition requires that, if we consider the mixture of each of any two lotteries L and L' with another lottery L'', the preference ordering on the two resulting lotteries is independent of the particular common lottery L''. That is, for any L, L' and L'', and any $\lambda \in (0, 1)$,

$$L \succsim L' \Leftrightarrow \lambda L + (1 - \lambda)L'' \succsim \lambda L' + (1 - \lambda)L''. \tag{2.1}$$

One may note that an equivalent assumption in the standard choice problem of consumer theory would be very restrictive, which is why it is seldom made in that context. Here, however, the independence assumption is quite natural because of a fundamental feature of decision problems under uncertainty: consequences are mutually exclusive.[2]

The independence axiom, coupled with the other standard rational choice assumptions, has the remarkable implication that there exists a utility function defined over consequences, $U : C \to \mathbb{R}$, such that

$$L \succsim L' \Leftrightarrow \sum_{i=1}^{N} \ell_i U(c_i) \geqslant \sum_{i=1}^{N} \ell_i' U(c_i), \tag{2.2}$$

where again, ℓ_i is the probability that consequence c_i will attain under L and ℓ_i' is the probability that consequence c_i will attain under L'. In other words, with the independence axiom, the utility function over lotteries can always be represented as the mathematical expectation of a utility function defined over consequences, that is $V(L) = \mathrm{E}[U(c)]$ where $\mathrm{E}[\cdot]$ is the mathematical expectation operator. As such, the utility function $V(L)$ is linear in probabilities. The function $U(c)$ is usually referred to as the von Neumann–Morgenstern (vNM) utility function.[3] This vNM utility function $U(c)$ is monotonically increasing and is cardinal in the sense that it is defined up to an increasing linear transformation [that is, if $U(c)$ represents the preference relation $\succsim$, then any $\widehat{U}(c) \equiv \alpha + \beta U(c)$, with $\beta > 0$, provides an equivalent representation of this relation]. When the outcomes of interest are described by continuous random variables with joint cumulative distribution function $F(c)$, the EU model implies that $V(F) = \int U(c)\,\mathrm{d}F(c)$. In conclusion, in the EU model the problem of selecting the action that induces the most preferred probability distribution reduces to that of maximizing the expected utility of outcomes.

Versions of the EU model more general than the one just discussed are available. Perhaps the most important is the EU model with subjective probability developed by Savage (1954).[4] In this framework one does not assume that the probabilities of various states of the world are objectively given. Rather, the existence of probabilities for the states of nature and of a vNM utility function for the consequences are both implied by a set of axioms. Prominent among these is the "sure-thing" axiom, roughly equivalent to the independence condition discussed earlier. A crucial element for this approach is that probabilities are inherently subjective, an idea pioneered by de Finetti (1931).

[2] Despite its theoretical appeal, the empirical validity of the independence axiom has been questioned, especially in light of the so-called Allais paradox [Allais (1953)].

[3] This convention recognizes these authors' pioneering contribution to the development of the EU model in [von Neumann and Morgenstern (1944)]. But others call $U(\cdot)$ the Bernoulli utility function, in recognition of Daniel Bernoulli's solution of the St. Petersburg paradox [Bernoulli (1738)], which anticipated some of the features of the EU model.

[4] Anscombe and Aumann (1963) provide an easier (albeit somewhat different) set-up within which one can derive Savage's subjective EU model.

2.2. Risk aversion

The EU model allows us to capture in a natural way the notion of risk aversion, which is a fundamental feature of the problem of choice under uncertainty. This notion is made precise when the consequences that matter to the decision maker are monetary outcomes, such that the vNM utility function is defined over wealth, say $U(w)$ where $w \in \mathbb{R}$ is realized wealth. In a very intuitive sense, a decision maker is said to be risk averse if, for every lottery $F(w)$, she will always prefer (at least weakly) the certain amount $\mathrm{E}[w]$ to the lottery $F(w)$ itself, i.e., $U[\int w\,\mathrm{d}F(w)] \geqslant \int U(w)\,\mathrm{d}F(w)$ [Arrow (1965), Pratt (1964)]. But by Jensen's inequality, this condition is equivalent to $U(w)$ being concave. Thus, concavity of the vNM utility function provides the fundamental characterization of risk aversion.

In many applied problems it is of interest to quantify risk aversion. For example, when can we say that an agent a is more risk averse than another agent b? Given the representation of risk aversion in terms of the concavity of $U(\cdot)$, then we can say that agent a is globally more risk averse than agent b if we can find an increasing concave function $g(\cdot)$ such that $U_a = g(U_b)$, where U_i denotes the utility function of agent i $(i = a, b)$. An interesting question, in this context, concerns how the degree of risk aversion of a given agent changes with the level of wealth. For this purpose, two measures of risk aversion that have become standard are the Arrow–Pratt coefficient of absolute risk aversion $A(w)$ and the Arrow–Pratt coefficient of relative risk aversion $R(w)$ [Arrow (1965), Pratt (1964)]. Because concavity of $U(w)$ is equivalent to risk aversion, the degree of concavity of $U(w)$, as captured for example by $U''(w)$, is a candidate to measure the degree of risk aversion. But because $U(w)$ is defined only up to an increasing linear transformation, we need to normalize by $U'(w) > 0$ to obtain a measure that is unique for a given preference ordering. Thus, the coefficient of absolute risk aversion is defined as $A(w) \equiv -U''(w)/U'(w)$.[5] As is apparent from its definition, absolute risk aversion is useful for comparing the attitude of an agent towards a given gamble at different levels of wealth. It seems natural to postulate that agents will become less averse to a given gamble as their wealth increases. This is the notion of decreasing absolute risk aversion (DARA), i.e., $A(w)$ is a decreasing function of w [when $A(w)$ is merely nonincreasing in w, the notion is labeled nonincreasing absolute risk aversion (NIARA)]. As we shall see, most comparative statics results of optimal choice under uncertainty rely on this condition.

Sometimes, however, it is interesting to inquire about the attitude of risk-averse decision makers towards gambles that are expressed as a fraction of their wealth. This type of risk preference is captured by the coefficient of relative risk aversion $R(w) \equiv wA(w)$. Unlike the case of absolute risk aversion, there are no compelling *a priori* reasons for

[5] Note that $A(w)$ can also be used to compare the risk aversion of two agents. If $A_a(w)$ and $A_b(w)$ are the coefficients derived from the vNM utility functions U_a and U_b, respectively, then agent a is more risk averse than agent b if $A_a(w) \geqslant A_b(w)$ for all w. This characterization is equivalent to that given earlier in terms of U_a being an increasing concave transformation of U_b.

any particular behavior of $R(w)$ with respect to w. An assumption that is sometimes invoked is that of nonincreasing relative risk aversion (NIRRA), implying that an agent should not become more averse to a gamble expressed as a fixed percentage of her wealth as the level of wealth increases.[6]

Of some interest for applied analysis are utility functions for which $A(w)$ and $R(w)$ are constant. The constant absolute risk aversion (CARA) utility function is given by $U(w) = -e^{-\lambda w}$, where λ is the (constant) coefficient of absolute risk aversion. The constant relative risk aversion (CRRA) utility function is given by $U(w) = (w^{1-\rho})/(1-\rho)$ if $\rho \neq 1$, and by $U(w) = \log(w)$ if $\rho = 1$, where ρ is the (constant) coefficient of relative risk aversion.[7]

2.3. Ranking distributions

As discussed, the choice problem under uncertainty can be thought of as a choice among distributions (lotteries), with risk-averse agents preferring distributions that are "less risky". But how can we rank distributions according to their riskiness? Earlier contributions tried to provide such ranking based on a univariate measure of variability, such as the variance or standard deviation [for example, the portfolio theory of Markowitz (1952) and Tobin (1958) relied on a *mean-standard deviation* approach]. But it was soon determined that, for arbitrary distributions, such ranking is always consistent with EU only if the vNM utility function is quadratic. Because of the restrictiveness of this condition, a more general approach has been worked out in what are known as the *stochastic dominance* conditions [Hadar and Russell (1969), Hanoch and Levy (1969), Rothschild and Stiglitz (1970)].

A distribution $F(w)$ is said to *first-order stochastically dominate* (FSD) another distribution $G(w)$ if, for every nondecreasing function $U(\cdot)$, we have

$$\int_{-\infty}^{\infty} U(w)\,\mathrm{d}F(w) \geqslant \int_{-\infty}^{\infty} U(w)\,\mathrm{d}G(w). \tag{2.3}$$

It can be shown that under FSD one must have $F(w) \leqslant G(w)$ for all w, a condition that provides an operational way of implementing FSD. Thus, this condition captures the idea that more is better in the sense that any agents for which w is a "good" should prefer $F(w)$ to $G(w)$. More to the point of choosing between distributions based on their riskiness, $F(w)$ is said to *second-order stochastically dominate* (SSD) another distribution $G(w)$ if the condition in (2.3) holds for every increasing and concave function $U(\cdot)$

[6] Arrow (1965) suggests that the value of $R(w)$ should hover around 1 and, if anything, should be increasing in w. His arguments are predicated on the requirement that the utility function be bounded, a condition that allows EU to escape a modified St. Petersburg paradox [Menger (1934)]. The relevance of these boundedness arguments for the behavior of $R(w)$, however, depends on $U(\cdot)$ being defined on the domain $(0, +\infty)$, a requirement that can be safely dropped in most applications.

[7] Note that, whereas CARA utility can be defined on $(-\infty, +\infty)$, CRRA utility is at most defined on $(0, +\infty)$. CARA and CRRA are special cases of the Hyperbolic Absolute Risk Aversion utility function.

[such that any risk averter will prefer $F(w)$ to $G(w)$]. It can be shown that in such a case one has

$$\int_{-\infty}^{w} [F(t) - G(t)]\,dt \leqslant 0 \tag{2.4}$$

for every w. Thus, (2.4) provides an operational characterization of SSD that can be used to compare distributions. A closely related notion is that of a *mean-preserving spread* [Rothschild and Stiglitz (1970)], which consists of taking probability mass away from a closed interval and allocating it outside that interval so that the mean of the distribution is unchanged. It turns out that, if a distribution function $G(\cdot)$ can be obtained from $F(\cdot)$ by a sequence of such mean-preserving spreads, then $F(\cdot)$ SSD the distribution $G(\cdot)$. Thus, when $F(w)$ and $G(w)$ have the same mean, the notion of a mean-preserving spread is equivalent to that of second-order stochastic dominance.

One should note that FSD and SSD produce only partial ordering of probability distributions. It is quite possible for any two distributions that neither one stochastically dominates the other, so that we cannot know for sure which one would be preferred by a particular risk-averse agent. Still, stochastic dominance and mean-preserving spreads give a precise characterization of what it means to have an increase in risk, and these conditions have proved to be extremely useful in analyzing the economic impact of changes in risk [Rothschild and Stiglitz (1971)].

When the distributions being compared are restricted to belonging to a particular class, it turns out that the validity of ranking distributions based on their mean and standard deviation can be rescued. In particular, if all distributions being compared differ from one another by a location and scale parameter only [i.e., $G(w) = F(\mu + \sigma w)$, where μ and σ are the location and scale parameters, respectively], then, as Meyer (1987) has shown, the mean-standard deviation ordering of distributions is quite general, in the sense that it is equivalent (for this class of distributions) to second-order stochastic dominance ordering.[8] The location-scale condition is restrictive (for example, it requires that an increase in variance occurs if and only if a mean-preserving spread occurs). Nonetheless, this condition applies to a number of interesting economic problems by the very definition of the problems themselves (for example, the theory of the competitive firm under price uncertainty) and also has some expositional value as discussed by Meyer (1987).[9]

3. The agricultural producer uuder uncertaiuty aud risk aversiou

The decision environment of agricultural producers is generally multifaceted and complex. Many distinct sources of risk may exist, and many discretionary actions may be

[8] As argued by Sinn (1989), there seem to exist earlier statements of this result.

[9] In any case, it should be clear that this result does not establish equivalence between EU and a linear mean-variance objective function, a criterion used in many agricultural economics applications.

available to the decision maker. Decisions and realizations of randomness may occur at several points in time. Further, actions may influence the distributions of yet-to-be realized random variables, while the realizations of random variables may alter the consequences of subsequent actions. To represent such an intricate network of interactions is analytically very difficult, but insights are possible by focusing on simpler stylized models. Thus, in the analysis that follows we start with an exceedingly simple model, and then gradually increase the complexity of the decision environment that we study. But first, an outline of model specifications that have the most relevance to agricultural decision making under uncertainty is in order.

3.1. Modeling price and production uncertainty

As outlined earlier, the main risks that a typical farmer faces are due to the fact that output prices are not known with certainty when production decisions are made and that the production process contains inherent sources of uncertainty (i.e., the relevant technology is stochastic). It is important, therefore, to understand how these fundamental sources of risk affect production decisions.

To capture the essence of price risk for competitive producers, consider the problem of choosing output q to maximize $\mathrm{E}[U(w_0 + \tilde{\pi})]$, where w_0 is the initial wealth and profit $\tilde{\pi}$ is random due to price uncertainty, that is,

$$\tilde{\pi} = \tilde{p}q - C(q, r) - K, \tag{3.1}$$

where $\tilde{p}$ denotes output price, $C(q, r)$ is the (variable) cost function (conditional on the vector of input prices r), and K represents fixed costs.[10] This is essentially the model considered by Sandmo (1971), among others. Note that, because there is no production uncertainty in this model, the technology of production has been conveniently represented by the cost function $C(q, r)$ so that the relevant choice problem can be couched as a single-variable unconstrained maximization problem.

When the production function is stochastic, it is clear that a standard cost function cannot represent the production technology [Pope and Chavas (1994)]. Thus, for the pure production uncertainty case, the production problem is best represented as that of choosing the vector of inputs x to maximize $\mathrm{E}[U(w_0 + \tilde{\pi})]$, with random profit given by

$$\tilde{\pi} = pG(x; \tilde{e}) - rx - K, \tag{3.2}$$

where $G(x; \tilde{e})$ represents the stochastic production function by which realized output depends on the vector of inputs x and a vector of random variables $\tilde{e}$. The latter represents factors that are important for production but are typically outside the complete

[10] To emphasize and clarify what the source of uncertainty is in any particular model, the overstruck ˜ will often be used to denote a random variable.

control of the farmer (examples include weather conditions, pest infestations, and disease outbreaks). It is clear that, in general, the production uncertainty case is more difficult to handle than the pure price risk case. In particular, it is typically necessary to restrict one's attention to the special case where $\tilde{e}$ is a single random variable. Versions of this model have been studied by Pope and Kramer (1979) and MacMinn and Holtmann (1983), among others.

Because price and production uncertainty are both relevant to agricultural production, it seems that the relevant model should allow for both sources of risk. Essentially, this entails making price p a random variable in (3.2). Joint consideration of price and production risk turns out to be rather difficult. Some results can be obtained, however, if the production risk is multiplicative, an assumption that was systematically used by Newbery and Stiglitz (1981), by Innes (1990), and by Innes and Rausser (1989). Specifically, the production function is written as $\tilde{e}H(x)$, where $\tilde{e}$ is a non-negative random variable (without loss of generality, assume $\mathrm{E}[\tilde{e}] = 1$), and so one chooses input vector x to maximize $\mathrm{E}[U(w_0 + \tilde{\pi})]$ with random profit given by

$$\tilde{\pi} = \tilde{p}\tilde{e}H(x) - rx - K. \tag{3.3}$$

Obviously, if the analysis is restricted to the consideration of a single random variable $\tilde{\varepsilon} \equiv \tilde{p}\tilde{e}$, it is clear that this model is isomorphic to the pure price risk case. In fact, as noted by a number of authors [Pope and Chavas (1994), Lapan and Moschini (1994), O'Donnell and Woodland (1995)], in this case there exists a standard cost function conditional on expected output, say $C(\overline{q}, r)$ where $\overline{q}$ is expected output,[11] that is dual to the production technology. Hence, the decision problem under joint price and (multiplicative) production risk can also be expressed as a single-variable unconstrained optimization problem because random profit in (3.3) can be equivalently expressed as

$$\tilde{\pi} = \tilde{p}\tilde{e}\overline{q} - C(\overline{q}, r) - K. \tag{3.4}$$

Before proceeding, we may note some restrictive features of the models just outlined. First, the models are static. There are essentially only two dates: the date at which decisions are made and the date at which uncertainty is realized (in particular, all decisions here are made before the resolution of uncertainty). Second, we are considering only one output and, for the time being, we are ignoring the possibility of risk management strategies. Although some of these assumptions will be relaxed later, such simplifications are necessary to get insights into the basic features of the production problem under risk.

In this setting, the basic questions that one may want to ask are:

(i) How does the existence of uncertainty affect choice?

(ii) Given uncertainty, how does a change in an exogenous variable affect choice? and

[11] Hence, for any given vector x of inputs, $\overline{q} = H(x)$.

(iii) To what extent does the existence of uncertainty alter the nature of the optimization problem faced by the decision maker?

For three of the basic contexts that we have outlined above (pure price risk, pure production risk with only one random variable, and joint price and production risk with multiplicative production risk), the answers to these questions can be characterized in a unified framework.

3.2. Static models under risk neutrality

Section 2 presented some concepts concerning the effects of riskiness on the expected value of a function. The first- and second-derivatives of a function were found to be key in determining how shifts in a stochastic distribution affect the expected value of a function. Although the structure of risk preferences, as expressed by the utility function, is certainly of consequence in determining the effects of risk on choice, risk-neutral decision makers may also be influenced by risk. Consider an expected profit-maximizing producer who faces a profile of profit opportunities $z(a, \beta, \tilde{\varepsilon})$ where a is a vector of choices (actions) at the discretion of the producer, β is a vector of exogenous parameters, and $\tilde{\varepsilon}$ is a single random variable that follows the cumulative distribution function $F(\varepsilon)$. Without loss of generality, let $\varepsilon \in [0, 1]$. The producer's problem is to

$$\operatorname*{Max}_{a} \int_0^1 z(a, \beta, \varepsilon)\, \mathrm{d}F(\varepsilon), \tag{3.5}$$

which yields the vector of first-order conditions

$$\int_0^1 z_a(a, \beta, \varepsilon)\, \mathrm{d}F(\varepsilon) = 0, \quad \text{where } z_a(\cdot) \equiv \partial z(\cdot)/\partial a.$$

Assuming that the choice vector is a singleton, and given concavity of $z(a, \beta, \tilde{\varepsilon})$ in a, from the concepts of stochastic dominance discussed earlier it is clear that an FSD shift in $\tilde{\varepsilon}$ will increase optimal a if $z_{a\varepsilon}(a, \beta, \varepsilon) \geqslant 0\ \forall \varepsilon \in [0, 1]$, whereas an SSD shift will increase optimal choice if, for all $\varepsilon \in [0, 1]$, $z_{a\varepsilon}(a, \beta, \varepsilon) \geqslant 0$ and $z_{a\varepsilon\varepsilon}(a, \beta, \varepsilon) \leqslant 0$.

A specification of $z(a, \beta, \tilde{\varepsilon})$ which is of immediate interest is that of pure price risk as given by (3.1), where $a \equiv q$ and where the stochastic output price satisfies $\tilde{p} = \beta_1 + (\tilde{\varepsilon} - \bar{\varepsilon})\beta_2$ (here $\bar{\varepsilon} \equiv \mathrm{E}[\tilde{\varepsilon}]$). One may interpret $\beta_1 + (\tilde{\varepsilon} - \bar{\varepsilon})\beta_2$ as a location and scale family of stochastic output price distributions with mean price equal to $\beta_1 \equiv \bar{p} \geqslant 0$, and the price variation parameter equal to $\beta_2 \geqslant 0$. Then the first-order condition for expected profit maximization is $\bar{p} - C_q(q, r) = 0$, and only the mean of the stochastic price is of relevance in determining optimal choice.

The more general form, where $\tilde{\varepsilon}$ cannot be separated out in this manner, may arise when production is stochastic. Then, even if $z_{a\varepsilon}(\cdot) \geqslant 0$, an increase in $\bar{\varepsilon}$ does not necessarily imply an increase in optimal a. The stochastic shift in $\tilde{\varepsilon}$ must be of the FSD dominating type, and an increase in the mean of $\tilde{\varepsilon}$ is necessary but insufficient for such a shift to occur.

It is also interesting to note that, in this risk-neutral case, an increase in an exogenous variable, say β_i, will increase optimal choice if $z_{a\beta_i}(a, \beta, \varepsilon) \geqslant 0\ \forall \varepsilon \in [0, 1]$, regardless of the distribution of $\tilde{\varepsilon}$:

$$\frac{\mathrm{d}a}{\mathrm{d}\beta_i} = -\frac{\int_0^1 z_{a\beta_i}(a, \beta, \varepsilon)\,\mathrm{d}F(\varepsilon)}{\int_0^1 z_{aa}(a, \beta, \varepsilon)\,\mathrm{d}F(\varepsilon)} \geqslant 0. \tag{3.6}$$

3.3. *Static models under risk aversion*

Given the payoff $z(a, \beta, \tilde{\varepsilon})$, the objective of a risk-averse producer is written as

$$\underset{a}{\mathrm{Max}} \int_0^1 U\big[z(a, \beta, \varepsilon)\big]\,\mathrm{d}F(\varepsilon), \tag{3.7}$$

where $U(\cdot)$ is increasing and concave, profit $z(\cdot)$ is held to increase in ε, and the objective function is concave in a, i.e., $\Delta \equiv \mathrm{E}\{U_{zz}[\cdot][z_a(\cdot)]^2 + U_z[\cdot]z_{aa}(\cdot)\} < 0$. Aspects of this problem, such as requirements on the nature of the utility function and payoff function and on the nature of the stochastic shift such that a increases, have been considered in some detail by Meyer and Ormiston (1983, 1985) and Eeckhoudt and Hansen (1992), among others. The first-order condition is

$$\int_0^1 U_z\big[z(a, \beta, \varepsilon)\big] z_a(a, \beta, \varepsilon)\,\mathrm{d}F(\varepsilon) = 0, \tag{3.8}$$

with parameterized solution at the value $a^* = a[F(\varepsilon), \beta]$.

3.3.1. *Introduction of uncertainty*

To ascertain how uncertainty affects choice for a risk averter, we will follow Krause (1979) and Katz (1981) and compare the solution under uncertainty with the solution when uncertainty is removed by setting the random element equal to its mean (i.e., setting $\tilde{\varepsilon} = \bar{\varepsilon}$). When uncertainty is removed, risk preferences are irrelevant, and the optimal choice $\hat{a}$ satisfies $z_a(\hat{a}, \beta, \bar{\varepsilon}) = 0$. When uncertainty exists, on the other hand, then the first-order condition can be expressed as

$$\mathrm{Cov}\big[U_z(\cdot), z_a(\cdot)\big] + \mathrm{E}\big[U_z(\cdot)\big]\mathrm{E}\big[z_a(\cdot)\big] = 0. \tag{3.9}$$

If $z_{a\varepsilon}(\cdot) \geqslant 0$, then the fact that the expectation of the product of two negatively covarying variates is less than the product of the expectations, together with risk aversion, implies that the covariance term must be negative. Because marginal utility is positive, satisfaction of the first-order condition requires that $\mathrm{E}[z_a(\cdot)] \geqslant 0$ when $z_{a\varepsilon}(\cdot) \geqslant 0$. We wish to compare a^*, the solution under uncertainty, with $\hat{a}$. If $z_{a\varepsilon\varepsilon}(\cdot) \leqslant 0$, then Jensen's inequality implies $\mathrm{E}[z_a(\hat{a}, \beta, \tilde{\varepsilon})] \leqslant 0$. But we know that $\mathrm{E}[z_a(a^*, \beta, \tilde{\varepsilon})] \geqslant 0$

given $z_{a\varepsilon}(\cdot) \geqslant 0$, and it follows that $\mathrm{E}[z_a(a^*, \beta, \tilde{\varepsilon})] - \mathrm{E}[z_a(\hat{a}, \beta, \tilde{\varepsilon})] \geqslant 0$. Because the only difference between the two expectations is the evaluation of a, and because $z_a(\cdot)$ is decreasing in a, then $a^* < \hat{a}$ [Krause (1979)]. The reduction in optimal a arises for two reasons. First, even for a risk-neutral producer, the existence of uncertainty reduces input use because it decreases the expected marginal value of the input, $\mathrm{E}[z_a(\cdot)]$. Second, risk aversion means that the increase in utility associated with an increase in $\tilde{\varepsilon}$ from $\bar{\varepsilon}$ is (in absolute value) lower than the decrease in utility associated with a decrease of the same magnitude in $\tilde{\varepsilon}$ from $\bar{\varepsilon}$. Because $z_{a\varepsilon}(\cdot) \geqslant 0$ (that is, an increase in a renders the payoff function more sensitive to the source of risk), the risk-averse producer will reduce sensitivity by decreasing a.

For an expected utility maximizer with payoff (3.1) (i.e., a competitive producer under price uncertainty only), it is clear that $z_{a\varepsilon}(\cdot) = 1 \geqslant 0$ and $z_{a\varepsilon\varepsilon}(\cdot) = 0 \leqslant 0$, so that the existence of price uncertainty reduces production. For payoff (3.2) (i.e., a competitive producer with stochastic production), $G_{a\varepsilon}(\cdot) \geqslant 0$ and $G_{a\varepsilon\varepsilon}(\cdot) \leqslant 0$ are sufficient conditions to sign the impact of introducing uncertainty. For a detailed analysis of input choice under stochastic production for risk-averse agents see Ramaswami (1992), who established requirements on an input-conditioned distribution function for a risk averter to choose less, or more, than an expected profit maximizer. A parallel analysis of Equation (3.9) shows that when $z_{a\varepsilon}(\cdot) \leqslant 0$ and $z_{a\varepsilon\varepsilon}(\cdot) \geqslant 0$, then risk aversion implies $a^* \geqslant \hat{a}$. The price uncertainty payoff [Equation (3.1)] never conforms to $z_{a\varepsilon}(\cdot) \leqslant 0$, but the production uncertainty model may. Thus, we see that the impact of the existence of uncertainty on optimal choice by a risk averter depends upon second and third cross-derivatives of the payoff function.

3.3.2. Marginal changes in environment

We now look at marginal changes in the decision environment, as represented by an increase in β. Intuitively, we know that the conditions required to identify the effects of these marginal changes are likely to be more stringent than those required to sign the effects of introducing uncertainty. Following Ormiston (1992), we differentiate Equation (3.8) partially with respect to a and β to obtain

$$\frac{\mathrm{d}a^*}{\mathrm{d}\beta} = \frac{1}{\Delta}\int_0^1 A[z]z_\beta(\cdot)U_z[\cdot]z_a(\cdot)\,\mathrm{d}F(\varepsilon) - \frac{1}{\Delta}\int_0^1 U_z[\cdot]z_{a\beta}(\cdot)\,\mathrm{d}F(\varepsilon), \tag{3.10}$$

where $A[\cdot] = -U_{zz}[\cdot]/U_z[\cdot]$is the absolute risk-aversion function defined earlier. Now we can partition the effect of β on a in three, which we will call (A) the wealth impact, (B) the insurance impact, and (C) the coupling impact [Hennessy (1998)]. The coupling impact is represented by the expression $-\int_0^1 U_z[\cdot]z_{a\beta}(\cdot)\,\mathrm{d}F(\varepsilon)/\Delta$ in (3.10) and has the sign of $z_{a\beta}(\cdot)$ if this term is uniform in sign. If β acts to increase the marginal effect of a on payoff $z(\cdot)$, then it will increase the producer's disposition to use a. For the price uncertainty case of (3.1) with $\tilde{p} = \beta_1 + (\tilde{\varepsilon} - \bar{\varepsilon})\beta_2$, we have $z_{a\beta_1}(\cdot) = 1$. For the

production uncertainty case of (3.2), where p is a nonstochastic shift variable, we have $z_{ap}(\cdot) = G_a(\cdot) > 0$.

Many agricultural support policies are constructed with the specific intent of having or not having a coupling effect. A price subsidy on an exogenous, institutional output or input quantity is decoupled in the sense that $z_{a\beta}(\cdot) = 0$, whereas with a true price subsidy the actual quantity is coupled. As an illustration, a modification of specification (3.1) is

$$z(a, \beta, \tilde{\varepsilon}) = \left[\beta_1 + (\tilde{\varepsilon} - \bar{\varepsilon})\beta_2\right]a - C(a, r) - K + \beta_3 G\left(a^0\right),$$

where $G(a^0)$ is some exogenous institutional reference production level. Here, $z_{a\beta_1}(\cdot) \geqslant 0$, but $z_{a\beta_3}(\cdot) = 0$. However, $z_{a\beta_2}(\cdot) = \tilde{\varepsilon} - \bar{\varepsilon}$ in this case, and this coupling effect does not have a uniform sign.

Effects (A) and (B) are intertwined in the first term on the right-hand side of (3.10). Let $J(\cdot, \varepsilon) \equiv A[\cdot]z_\beta(\cdot)$, so the expression is $Q \equiv \int_0^1 J(\cdot, \varepsilon)U_z[\cdot]z_a(\cdot)\,\mathrm{d}F(\varepsilon)/\Delta$. Integrating by parts yields

$$\begin{aligned} Q &= \frac{1}{\Delta}\left[J(\cdot, \nu)\int_0^\nu U_z[\cdot]z_a(\cdot)\,\mathrm{d}F(\varepsilon)\Big|_{\nu=0}^{\nu=1} - \int_0^1\int_0^\nu U_z[\cdot]z_a(\cdot)\,\mathrm{d}F(\varepsilon)\frac{\mathrm{d}J(\cdot, \nu)}{\mathrm{d}\nu}\,\mathrm{d}\nu\right], \\ &= -\frac{1}{\Delta}\int_0^1\int_0^\nu U_z[\cdot]z_a(\cdot)\,\mathrm{d}F(\varepsilon)\frac{\mathrm{d}J(\cdot, \nu)}{\mathrm{d}\nu}\,\mathrm{d}\nu, \end{aligned} \tag{3.11}$$

where ν is used as the dummy variable of integration for the variable ε. To identify effects (A) and (B) note that, if $z_{a\varepsilon}(\cdot) \geqslant 0$, the first-order condition (3.8) implies that the expression $\int_0^\nu U_z[\cdot]z_a(\cdot)\,\mathrm{d}F(\varepsilon)$ is never positive because of the positivity of marginal utility and because $z_a(\cdot)$ is negative at low ε and increases to be positive at high ε. Therefore, given $\Delta < 0$, Q is positive if $\mathrm{d}J(\cdot, \nu)/\mathrm{d}\nu \leqslant 0$. Differentiate to obtain $\mathrm{d}J(\cdot, \nu)/\mathrm{d}\nu = z_\beta(\cdot)A_z[\cdot]z_\varepsilon(\cdot) + A[\cdot]z_{\beta\varepsilon}(\cdot)$. The first part of this expression may be called the wealth effect (A) because its negativity depends upon the NIARA property and the sign of $z_\beta(\cdot)$ (recall that $z_\varepsilon(\cdot) \geqslant 0$). All other things equal, if β shifts the distribution of payoffs rightward ($z_\beta(\cdot) \geqslant 0$), as would be the case with a reduction in fixed costs K in payoff specifications (3.1) or (3.2), and if preferences are NIARA ($A_z[\cdot] \leqslant 0$), then a increases. When $\tilde{p} = \beta_1 + (\tilde{\varepsilon} - \bar{\varepsilon})\beta_2$, then $z_{\beta_1}(\cdot) \geqslant 0$ for specification (3.1). Because $z_{a\beta_1}(\cdot) \geqslant 0$, both coupling and wealth effects act to increase optimal a, and this is the Sandmo (1971) result that NIARA is sufficient for a shift in mean price to increase production. Notice that because $z_{\varepsilon\beta_1}(\cdot) = 0$, the second part of $\mathrm{d}J(\cdot)/\mathrm{d}\nu$ may be ignored. Whereas β_1 has both wealth and coupling effects, it is easy to describe a wealth effect that does not also involve coupling. Setting $z(a, \beta, \tilde{\varepsilon}) = [\beta_1 + (\tilde{\varepsilon} - \bar{\varepsilon})\beta_2]a - C(a, r) - K + \beta_3 G(a_0)$, an increase in β_3 or a decrease in K induces an increase in optimal a under NIARA. Coupling may also occur without wealth effects, although this case is somewhat more difficult to show.

The second part, $A[\cdot]z_{\beta\varepsilon}(\cdot)$, is the insurance effect (B). If the favorable exogenous shift acts to stabilize income, that is if $z_{\beta\varepsilon}(\cdot) \leqslant 0$ or β advances less fortunate states of

the environment by more than it advances more fortunate states, then optimal a tends to increase. This would occur in specification (3.2) if $\beta = p$ and $G_{p\varepsilon}(\cdot) \leqslant 0$. In the case of an insurance contract on the source of uncertainty, say $M(\beta, \tilde{\varepsilon})$, the payoff is $pG(a, \tilde{\varepsilon}) - wa - K + M(\beta, \tilde{\varepsilon})$ and the insurance contract decreases risk if $M_{\beta\varepsilon}(\cdot) \leqslant 0$ while $pG_{\varepsilon}(\cdot) + M_{\varepsilon}(\cdot) \geqslant 0$. The similarity of wealth (i.e., risk aversion) and insurance effects has been discussed in detail by Jewitt (1987).

Because of the price uncertainty inherent in agricultural production environments, the effect of an increase in β_2 for the specification (3.1), where $\tilde{p} = \beta_1 + (\tilde{\varepsilon} - \bar{\varepsilon})\beta_2$, is of particular importance. From $z_{\varepsilon\beta_2} = 1$, it can be seen that the β_2 parameter has a negative insurance effect. It has already been concluded, however, that the coupling effect of β_2, that is $z_{a\beta_2}(\cdot)$, does not have a uniform sign. Thus, although it may be intuitive to expect that an increase in β_2 would decrease optimal a, to determine that requires more work in addition to the NIARA assumption [Batra and Ullah (1974), Ishii (1977)]. Since changing the parameter $\beta_2 \geqslant 0$ in this setting does not cover the set of all Rothschild and Stiglitz mean-preserving spreads, the above results do not demonstrate that all mean-preserving spreads of price decrease the optimal choice for the model in (3.1). Whereas Meyer and Ormiston (1989), Ormiston (1992), and Gollier (1995), among others, have made advances toward identifying precisely the set of spreads that act to decrease production for NIARA and various conditions on the payoff function, this problem has not yet been completely solved.[12]

3.3.3. Uncertainty and cost minimization

It is well known that profit maximization is predicated upon satisfaction of the cost minimization assumption. Does cost minimization continue to hold under risk, when the objective is expected utility maximization? It turns out that the answer is yes, provided that "cost minimization" is suitably defined. Consider the competitive firm where the input vector x is chosen to maximize $\mathrm{E}[U(w_0 + \tilde{\pi})]$, where $\tilde{\pi} = R(x, \tilde{\varepsilon}) - rx$. Here $R(x, \tilde{\varepsilon})$ is a revenue profile (that can accommodate both price and/or production uncertainty) and $\tilde{\varepsilon}$ denotes the source of revenue uncertainty. Pope and Chavas (1994) show that, if the revenue profile satisfies the restriction $R(x, \tilde{\varepsilon}) = K(\psi(x), \tilde{\varepsilon})$, where $\psi(x)$ is (possibly) vector-valued, then the relevant cost function can be written as $C(q^{\psi}, r)$, where q^{ψ} is the vector of conditioning values corresponding to the functions $\psi(x)$. Hence, technical efficiency is satisfied in the sense that the EU maximizing choice of x is consistent with the cost minimizing means of obtaining some (vector) level of $\psi(x)$. The simplest special case arises with multiplicative production risk, when $R(x, \tilde{\varepsilon}) = H(x)\tilde{\varepsilon}$. As anticipated in Section 3.1, in such a case the cost function is written as $C(\overline{q}, r)$, where $\overline{q}$ is

[12] The conclusions drawn thus far are, of course, only relevant for the given context. Noting that peasants in less developed countries often consume a significant fraction of their own production, Finkelshtain and Chalfant (1991) concluded that production and consumption decisions cannot be modeled separately for these agents. Their generalization of the Sandmo model suggests that production may plausibly increase under an increase in price uncertainty.

expected output. Thus, the relevant cost function for this special case is rather standard, with the expected output level playing the role of a deterministic output level under certainty. More generally, however, a vector of conditioning values will be needed. For example, if there is no price risk but the production function has the stochastic form suggested by Just and Pope (1978) (to be discussed further in Section 4.2), then revenue is written as $R = pM(x) + p[V(x)]^{1/2}\tilde{e}$ with $\mathrm{E}[\tilde{e}] = 0$. It follows that the EU-consistent cost function here can be written as $C(\overline{q}, \sigma^2, r)$, where $\overline{q}$ is a level of expected output [corresponding to the function $M(x)$] and σ^2 is a level of output variance [corresponding to the function $V(x)$].

That cost minimization always holds for EU maximizers, even when the revenue profile does not satisfy the restriction invoked by Pope and Chavas (1994), is shown by Chambers and Quiggin (1998). Their approach is best illustrated for the production uncertainty case in which the random variable $\tilde{e}$ takes on a finite number (say N) of values. Given the stochastic production function $G(x, \tilde{e})$, then realized output for any given realization of the random variable (e_i, say) is $q_i = G(x, e_i)$. If ℓ_i denotes the probability of e_i occurring, then the producer's EU problem is

$$\operatorname*{Max}_{x} \sum_{i=1}^{N} \ell_i U\big(pG(x, e_i) - rx\big). \tag{3.12}$$

Now define a cost function $C(q_1, \ldots, q_N, r)$ as

$$C(q_1, q_2, \ldots, q_N, r) \equiv \operatorname*{Min}_{x}\big\{rx\colon\ q_i \leqslant G(x, e_i),\ \forall i = 1, 2, \ldots, N\big\}. \tag{3.13}$$

One may note the formal similarities of $C(q_1, \ldots, q_N, r)$ with a standard multioutput cost function, although the interpretation here is rather different. At any rate, it follows that the producer's EU maximization problem can be equivalently expressed as

$$\operatorname*{Max}_{q_1, q_2, \ldots, q_N} \sum_{i=1}^{N} \ell_i U\big(pq_i - C(q_1, q_2, \ldots, q_N, r)\big). \tag{3.14}$$

Thus, it is clear that EU maximizers do minimize costs, in some sense.

3.4. Dynamics and flexibility under uncertainty

A consideration of decision making under risk is not complete without discussion of the interactions between risk and time. Although suppressed in the two dates (one period) models discussed above (i.e., action at time 0 and realization at time 1), the fact is that time and uncertainty are intertwined because information sets become more complete as time passes. To illustrate, we consider a simple extension of the price uncertainty case of model (3.1). Specifically, let $a \equiv (x_1, x_2)$ such that $z(a, \beta, \tilde{\varepsilon})$ is of form $\tilde{\varepsilon}R(x_1, x_2) - r_1x_1 - r_2x_2$ where $\tilde{\varepsilon}$ represents stochastic output price, and assume that x_1 is chosen

before the realization of $\tilde{\varepsilon}$, whereas x_2 is chosen after $\tilde{\varepsilon}$ is observed. Following Hartman (1976), the problem may be posed as

$$\underset{x_1}{\text{Max}} \int_0^1 \underset{x_2}{\text{Max}}\left[\varepsilon R(x_1, x_2) - r_2 x_2\right] \mathrm{d}F(\varepsilon) - r_1 x_1. \tag{3.15}$$

Applying backward induction, the second-stage problem is solved first. The first-order condition is $\varepsilon R_{x_2}(x_1, x_2) = r_2$, where x_1 and ε are now predetermined. Assuming strict concavity of $R(\cdot)$ in x_2, the first-order condition is solved to yield $x_2^* = S(x_1, r_2, \varepsilon)$. Given this short-run demand function for x_2, the producer problem reduces to

$$\underset{x_1}{\text{Max}} \int_0^1 \left[\varepsilon R\big(x_1, S(x_1, r_2, \varepsilon)\big) - r_2 S(x_1, r_2, \varepsilon)\right] \mathrm{d}F(\varepsilon) - r_1 x_1. \tag{3.16}$$

Defining $L(x_1, r_2, \varepsilon) \equiv \varepsilon R(x_1, S(x_1, r_2, \varepsilon)) - r_2 S(x_1, r_2, \varepsilon)$, the envelope theorem gives the first-order condition for the first-stage problem (choosing x_1) as

$$\int_0^1 L_{x_1}(x_1, r_2, \varepsilon) \, \mathrm{d}F(\varepsilon) - r_1 = 0. \tag{3.17}$$

Now, the Rothschild and Stiglitz mean-preserving spread condition implies that optimum x_1 increases with such a spread if $L_{x_1\varepsilon\varepsilon}(x_1, r_2, \varepsilon) \geqslant 0$. Setting optimum output as $G^*(x_1, r_2, \varepsilon)$, the envelope theorem can be used to show that this is the same as requiring $G^*_{x_1\varepsilon}(x_1, r_2, \varepsilon) \geqslant 0$. Further analysis reveals that this condition is equivalent to the requirement that $\partial[R_{x_1x_2}(\cdot)/R_{x_2x_2}(\cdot)]/\partial x_2 \leqslant 0$. Thus, when ex-post flexibility exists, the effects of uncertainty depend upon relationships between third derivatives of the production technology. In general, although the impact of a mean-preserving spread in ε on the distribution of x_2^* depends upon the sign of $\partial^2 S(\cdot)/\partial\varepsilon^2$, the impact on x_1 is less readily signed and the effect on mean $R(\cdot)$ is yet more difficult to sign. Obviously, the analysis becomes even more involved when decision makers are assumed to be risk averse.

A second set of problems, called real option problems because of structural analogies with financial options, arise from the interactions between time and uncertainty in long-term investment decisions when there are sunk costs or irreversible actions. Consider a decision in 1999 to invest in precision farming education and equipment. At that time it was not yet clear whether the technology was worth adopting. The decision maker may invest early in the hope that the technology will turn out to be profitable. But the investment may turn out to be unprofitable, so there is also an incentive to defer the decision for a year, say, to learn more about the technology in the intervening period. But deferment will mean losing a year of additional profits if the technology turns out to be profitable. Similar sunk cost and information problems may arise in a number of other farm production decisions. Although real option problems such as these can be addressed by rigorous stochastic neoclassical models [e.g., Chavas (1994) or Feinerman et al. (1990)]

or by standard optimal control approaches [Rausser and Hochman (1979)], the more structured contingent claims approach popularized by Dixit and Pindyck (1994) has assumed prominence because it lends itself to empirical and theoretical analysis.

A stylized continuous-time variant of dynamic programming, real option theory connects time and uncertainty by modeling a source of randomness as a stochastic process evolving over time. Some such processes give rise to differential equation relationships between the distribution, time, and the flow of rewards. These relationships can be solved to give a decision-conditioned expected present value, and this expected present value is then optimized over the choice set. The choice set may involve deciding to invest now or to wait, or deciding how much to invest. Marcus and Modest (1984) studied optimal decisions for producers facing price and yield uncertainty and using futures markets, whereas Turvey (1992b) used the approach to study agricultural support policies in Canada. Purvis et al. (1995) adopted the framework to explain Texas dairy industry technology adoption decisions under cost and regulatory uncertainty, and found that the expected rate of return on the proposed investment might have to be double the threshold identified by a nonstochastic analysis for the decision to be attractive. The approach also provides a simple way of studying adjustment costs. For example, Leahy (1993) studied shutdown and startup costs for a competitive firm facing random prices.

4. Selected empirical issues

Our cursory review thus far has focused on analytical methods and theoretical analyses. But considerable empirical research in agricultural economics has been done to test, quantify, and otherwise put to use a number of features of risk models. In this section we will look, in some detail, at a number of contributions that have had a primarily empirical bent.

4.1. Identifying risk preferences

In an early empirical study of agricultural decision making under risk, Lin, Dean and Moore (1974) elicited preferences over hypothetical lotteries from managers of six large California farms. Using quadratic programming methods, they estimated the mean-variance frontier available to the farmer. They then compared the farm plans suggested by the elicited preference structure with plans suggested by the expected profit maximization rule, with plans suggested by lexicographic preference structures, and with the actual implemented plans. They found that, although no stylized preference structure was clearly a superior fit, for each of the six farms the EU framework performed at least as well as the other paradigms. For Nepalese rice farmers, Hamal and Anderson (1982) also used hypothetical lotteries and found evidence in support of DARA. The analysis was less conclusive concerning the slope of relative risk aversion.

Dillon and Scandizzo (1978) modified the approach of Lin, Dean and Moore (1974) by eliciting preferences from a relatively large number of subsistence farmers and share-

croppers in northeastern Brazil. Risk attitudes were imputed from choices between hypothetical lotteries that realistically reflected the farm payoffs faced by these decision makers. Unlike the study by Lin, Dean and Moore, however, the hypothetical decisions were not validated through comparison with actual decisions. The lotteries posed were of two types, those in which the family subsistence requirement was covered but surplus income was at risk, and those in which the subsistence requirement was also at risk. Hypothetical returns were adjusted until certainty equivalence between lottery comparisons was established. The replies were then fitted to three decision criteria: mean-standard deviation, mean-variance, and CARA expected utility objective functions. As expected, both farmers and sharecroppers tended to be more risk averse when subsistence income was at risk. Surprisingly, smallholders tended to be more risk averse than sharecroppers. Dillon and Scandizzo (1978) found less clear evidence about the impact of socioeconomic factors on risk attitudes. Perhaps the most interesting indication was that, even within seemingly homogeneous groups, a wide dispersion of risk preferences appeared to exist.

Taking an econometric approach, Moscardi and de Janvry (1977) estimated a Cobb–Douglas production function for corn with data from small Mexican subsistence farms. Using a safety-first framework, they imputed a measure of risk aversion from the divergence between actual fertilizing decisions and optimal decisions under risk neutrality. They found evidence of considerable risk aversion, and they also suggested that risk attitudes might be functions of socioeconomic variables (such as family size and age of operator) that may evolve over time. Brink and McCarl (1978) also estimated risk attitudes as a residual that rationalizes observed choices relative to "optimal" ones as predicted by a mathematical programming model (relying on a linear mean-standard deviation objective function). Thirty-eight Midwestern crop producers at a Purdue University decision analysis workshop listed their resources and identified their preferred crop acreage allocation plan. The risk parameter giving a plan deemed closest to the announced plan was assumed to represent risk preferences. The analysis concluded that risk aversion seemed to be low. Measuring risk essentially as a residual, however, is an obvious limitation of these studies (because such a procedure ignores other potential reasons for observed decisions to depart from the model's optimal decisions).

Because of the limitations of inferring risk from observed production decisions, and because hypothetical payout surveys can give unstable results, Binswanger (1980) made real payments to peasant farmers in India. Outcomes were determined by tossing a dice, and the amount at risk varied from 0.5 rupees to 500 rupees (negative payout states were not considered). The 500 rupees payout amounted to about 2.3 percent of average household wealth, and corresponded in magnitude to substantial fertilization investments. (It was believed that some households were constrained by capital resources from fertilizing adequately.) Preliminary tests found that individuals tended to treat money gifted to them on the day of the experiment for the purpose of participating in the experiment as if it were their own. Preliminary results also suggested that once lotteries for low gambles had primed individuals to making lottery decisions about real money, then a hypothetical 500 rupees game appeared to give results that were statistically similar to

a real 500 rupees game. To conserve financial resources, the hypothetical 500 rupees game was used thereafter.

Capturing risk attitudes by the coefficient of partial risk aversion, it was found that subjects tended to become more risk averse as the size of the gamble increased.[13] Compared with the hypothetical scenario interviewing method, the imputed risk aversion coefficient was less dispersed when real money was involved. This would suggest that the interviewers may have had difficulty taking the interviews as seriously as they would real-world decisions. On the effects of socioeconomic characteristics, Binswanger (1980) found that wealthier, better-educated, and more progressive farmers tended to be less risk averse, as did those who had off-farm salaries. Prior luck in the game also tended to reduce the degree of risk aversion (only the luck regressor, however, had consistently high t statistics across all gamble sizes). Overall, Binswanger interpreted the results as being supportive of the hypothesis that it is resource and infrastructural constraints, such as access to information and credit, that induce caution among peasants rather than the hypothesis of innate conservatism.

In a different analysis of these Indian data, Binswanger (1981) considered the foundations of the EU framework and concluded that decision makers did not integrate possible outcomes from a gamble with pre-existing income, but rather treated them separately in their decision calculus. This conclusion is somewhat at variance with Binswanger's (1980) conclusion from pretest analysis that subjects treated gifted money as their own. The separation of gamble money from pre-existing wealth lends some support to Kahneman and Tversky's (1979) prospect theory approach to decision making. Failure of income integration has serious implications for modeling decisions, but has generally been ignored in the empirical literature. Binswanger also used inferences drawn from safety-first type models to identify inconsistencies with the data, and he concluded that the decision makers did not appear to act in a safety-first manner. Finally, Binswanger identified evidence in the data to support both DARA and decreasing relative risk aversion (DRRA) preferences.

Surveying work on risk preferences and risk management to that time (including work by Binswanger already cited), Young (1979) and Hazell (1982) raised concerns about all approaches. The direct elicitation (interview) method is reliable only to the extent that it captures the preference structure that would be used in real decisions, and evidence suggested that it might not do so. Experimental approaches might be too expensive to implement in developed countries.[14] Approaches based on observed supply and input demand behavior impute risk as the residual component explaining discrepancies between expected profit-maximizing solutions and actual decisions. But discrepancies may be due to other effects, such as imperfect information and heterogeneous resource

[13] The coefficient of partial risk aversion is defined as $-\pi U''(\pi + w_0)/U'(\pi + w_0)$, where w_0 is initial wealth and π is profit.

[14] Binswanger estimated that, were he to run his experiments in the United States, it would have cost $150,000 (circa 1978) rather than $2,500.

endowments. To the extent that such research had identified determinants of risk preferences, Young concluded that farmers in developing countries appeared to be more risk averse than those in developed countries, and he observed that this conclusion is consistent with DARA. But because the studies considered did not explicitly control for the availability and use of risk management institutions, which tend to be more widely available in developed countries, developed-country farmers may appear to be less risk averse than they actually are.

Returning to the task of econometrically estimating risk structures, Antle (1987) expressed the optimality conditions for EU maximizing choices in terms of a given individual's absolute risk aversion and downside risk aversion coefficients.[15] The Generalized Method of Moments (GMM) procedure was then applied to identify means, variances, and covariances of risk preference parameters based on data from the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) pertaining to one of the six Indian villages (Aurepalle) that had been considered by Binswanger (1980, 1981). Antle (1987) found a mean Arrow–Pratt index similar to that reported in Binswanger (1980). Dissatisfied that this approach required some, if only minimal, assumptions concerning the technology available, Antle (1989) developed a method that did not involve joint estimation with technology. Antle's view was that it would be better to estimate risk preference structures separately from technology rather than jointly. His concerns about a joint estimation arose mainly from problems involving the data required for the estimation of technology, and the discontent with alternative econometric approaches to joint estimation. The econometric methods applied again involved GMM estimation on data from the ICRISAT India village study. The means of the Arrow–Pratt and downside risk aversion indices were, as expected, similar to those estimated earlier.

Among other econometric estimations of risk attitudes, Myers (1989) assumed CRRA and joint lognormality of the distributions of output price and producer consumption, and developed a reduced-form rational expectations approach to testing for the aggregate level of relative risk aversion for U.S. producers who store crops. Annual data over the period 1945 to 1983 suggest a coefficient of relative risk aversion between 1.5 and 4.5 for corn and wheat storers, but the estimates for soybeans are implausible. Exploiting technical attributes of CRRA and of constant partial relative risk aversion (CPRRA),[16] Pope (1988) developed implications for optimal choices by individuals expressing such preferences. In Pope and Just (1991), these implications, together with implications for choice under CARA preferences, were tested on state-level Idaho potato acreage data. CARA and CPRRA hypotheses were rejected, but CRRA was not. Chavas and Holt (1990), studying U.S.-level corn and soybean acreage allocation decisions, also used the tests proposed by Pope (1988) and rejected both CRRA and CPRRA. Testing for the impact of wealth, proxied by an index of proprietor equity, on allocation decisions, they found evidence to reject CARA in favor of DARA.

[15] This downside risk aversion coefficient is defined as $U'''(\cdot)/U'(\cdot)$. Note that $U'''(\cdot) > 0$ is necessary for DARA. For the related, but distinct, coefficient of absolute prudence $(-U'''(\cdot)/U''(\cdot))$ see Kimball (1990).

[16] This means that $-\pi U''(\pi + w_0)/U'(\pi + w_0)$ is invariant to changes in π for the level of w_0 in question.

4.2. Estimating stochastic structures

As mentioned earlier, production risk is an essential feature of agriculture, and estimation of such stochastic production structures has obvious immediate interest for farm management as well as to address agricultural policy issues. For example, production uncertainty has implications for the implementation of crop insurance. Also, environmental externalities such as water contamination and ecosystem destruction may sometimes be traced back to the use of such agricultural inputs as nitrogen and pesticides; production uncertainty, together with risk aversion, may increase application of these inputs. Existing statistical procedures for studying relationships between stochastic distributions have tended to emphasize stochastically ordered comparisons, such as first- and second-degree dominance, between elements in a set of distributions. But economists, especially agricultural economists, are often interested in conditional relationships. To reconstruct nonparametric stochastic relationships between crop yield and input use would often require volumes of data beyond that usually available to analysts. Further, as the literature on the impacts of stochastic shifts on decisions has shown, the necessary and the sufficient conditions for a stochastic shift to have a determinate impact on the decisions of a meaningful class of decision makers are generally not among the simpler types of stochastic shifts.

The complexity of the decision environment is substantially reduced if one can treat technology as being nonrandom. If one is primarily concerned with price uncertainty, then it might be convenient to assume deterministic production. Thus, one can estimate the distribution of the realized random element without regard to the choices made. In other cases, however, it is not possible to simplify the decision environment in this way. Although random yield – the consequence of interactions between choices and random weather variables – can be measured, it would be more difficult to measure and aggregate in a meaningful manner the various dimensions of weather. In such a case, it is more convenient to estimate the input-conditioned distribution of yield. Although they do not lend themselves to estimating or testing for general production function relations, existing stochastic ordering methods can be useful in testing for the nature of and impacts of exogenous stochastic shifts in, say, the distribution of output price, and for studying discrete decisions such as the adoption of a new technology.

Although studies applying stochastic dominance methods to agricultural problems are numerous [e.g., Williams et al. (1993)], most of these studies compare point estimates of the distributions and do not consider sampling errors. Tolley and Pope (1988) developed a nonparametric permutation test to discern whether a second-order dominance relationship exists. More recently, Anderson (1996) used the nonparametric Pearson goodness-of-fit test on Canadian income distribution data over the years 1973 to 1989 to investigate, with levels of statistical confidence, whether first-, second-, and third-order stochastic dominance shifts occurred as time elapsed.

For input-conditioned output distributions, Just and Pope (1978) accounted for heteroskedasticity by developing a method of estimating a two-moment stochastic produc-

tion function by three-stage non linear least squares techniques. The function is of the form

$$\tilde{q} = M(x) + \left[V(x)\right]^{1/2}\tilde{\varepsilon}, \tag{4.1}$$

where q is output, $\mathrm{E}[\tilde{\varepsilon}] = 0$, $\mathrm{Var}[\tilde{\varepsilon}] = 1$, and x is a vector of input choices. The functions $M(x)$ and $V(x)$ determine the conditional mean and variance of q, respectively, and can be chosen to be sufficiently flexible to meet the needs of the analysis. Just and Pope (1979) applied their method to Day's (1965) corn and oats yield-fertilization data set, and found the results generally, but not totally, supportive of the hypothesis that an increase in fertilization increases the variance of output. Their readily estimable approach has proven to be popular in applied analyses. For example Traxler et al. (1995) used the approach in a study of the yield attributes of different wheat varieties in the Yaqui Valley (Mexico), and found that whereas earlier varietal research appeared to emphasize increasing mean yield, later research appeared biased toward reducing yield variance.

Suggesting that mean and variance may not be sufficient statistics to describe stochastic production, Antle and Goodger (1984) used an approach due to Antle (1983) to estimate an arbitrarily large number of input-conditioned moments for large-scale California milk production. They rejected the statistical hypothesis that input-conditioned mean and variance are sufficient statistics. An interesting simulation finding was that a CARA decision maker facing the estimated technology substantially increased dairy rations relative to a risk-neutral decision maker. This suggests that the marginal risk premium in Ramaswami (1992) may be negative on occasion.

Nelson and Preckel (1989) identified the need for a flexible approach to estimating parametric yield distributions when accommodating skewness is important. Gallagher (1987), among others, has observed negative skewness for crop yields. The Just–Pope approach is insufficiently flexible, whereas the Antle–Goodger method, which is nonparametric, may be inefficient. Finding inspiration in Day's (1965) suggestion that the beta distribution would likely fit most yield distributions quite well, Nelson and Preckel conditioned beta distribution parameters on input choices. The output density function is then

$$f(q \mid x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{(q - q^{\min})^{\alpha-1}(q^{\max} - q)^{\beta-1}}{(q^{\max} - q^{\min})^{\alpha+\beta-1}}, \tag{4.2}$$

where $\Gamma(\cdot)$ is the gamma function, output q is supported on the interval $[q^{\min}, q^{\max}]$, and the distribution parameters are conditional on inputs, i.e., $\alpha = \alpha(x)$ and $\beta = \beta(x)$. For field-level corn yields in five Iowa counties over the period 1961 to 1970, Nelson and Preckel set $q^{\min} = 0$, and let both $\alpha(x)$ and $\beta(x)$ be Cobb–Douglas functions of nitrogen, phosphorus, potassium, field slope, and soil clay content. Using a two-stage maximum likelihood method, they found that the marginal effects of nitrogen, phosphorus, and potassium on skewness, variance, and even mean were mixed in sign.

The maximum likelihood approach to estimating parameterized conditional densities has proven to be quite popular. A gamma distribution relationship between applied nitrogen levels and late spring soil nitrate levels has been used in Babcock and Blackmer (1992) to study the effects of information concerning spring soil nitrate levels on subsequent side-dressing and on expected profit; a beta distribution has been applied by Babcock and Hennessy (1996) to study input use in the presence of crop insurance. A different line of inquiry has sought to model the nonnormality of crop yield distributions by estimating transformations of the normal distribution. Taylor (1990) employed a hyperbolic trigonometric transformation to deviations from a linear yield trend estimation on corn, soybean, and wheat crops. Moss and Shonkwiler (1993) and Ramírez (1997) have extended this approach to accommodate stochastic yield trends and multivariate distributions, respectively. But the presumption that yields are not normally distributed has been called into question by Just and Weninger (1999), who criticize a number of features of statistical analyses implemented by previous studies and conclude that the empirical evidence against normality is weak.

Stochastic production has implications for the estimation of dual representations of production technologies. For example, as discussed in Section 2.3.3, when the production function is affected by multiplicative risk and producers maximize expected utility the relevant cost function is $C(\overline{q}, r)$, where $\overline{q}$ is expected output. When the stochastic production function is written more generally as $G(x, \tilde{e})$, the relevant cost function still has the structure $C(\overline{q}, r)$ if producers are risk neutral (they maximize expected profits).[17] Pope and Just (1996) call such a function the "ex ante cost function", and convincingly argue that a number of previous studies have resulted in inconsistent estimates of technological parameters because they have estimated a standard cost function $C(q, r)$ (conditional on realized output q) when in fact they should have been estimating $C(\overline{q}, r)$. Estimation of the ex ante cost function $C(\overline{q}, r)$is problematic, on the other hand, because it is conditional on expected output $\overline{q}$, which is not observable. The solution proposed by Pope and Just (1996) entails estimating $\overline{q}$ jointly with the structure of the ex ante cost function. The specific procedure that they suggest fails to achieve consistent estimation of technological parameters because it does not address the nonlinear errors-in-variables problem that typically arises in this context [Moschini (1999)]. But by exploiting the full implications of expected profit maximization, Moschini (1999) shows that it is possible to effectively remove the errors-in-variables problem and obtain consistent estimation of the ex ante cost function parameters.

4.3. Joint estimation of preferences and technology

Most research studies considered thus far have sought to identify risk preferences without estimating the source of randomness, or they have sought to estimate the source

[17] Of course, in such a case the parameters of the cost function $C(\overline{q}, r)$ may include parameters of the distribution of the random variable $\tilde{e}$.

of randomness without simultaneously estimating the risk preference structure. Those papers that have simultaneously identified risk preferences and the source of randomness [e.g., Moscardi and de Janvry (1977)) or Antle (1987)] have treated either one or both components in a rather elementary manner. Separating the estimation of the two structures is econometrically inefficient to the extent that a joint estimation imposes cross-estimation restrictions and accommodates error correlations. Using a Just–Pope technology with Cobb–Douglas mean and variance functions together with a CARA risk preference structure, cross-equation restrictions and a nonlinear three-stage least squares estimator, Love and Buccola (1991) applied a joint estimation for Iowa corn and soybean production. The data pertained to three of the five counties studied by Nelson and Preckel (1989). Love and Buccola found considerable variation in the estimated coefficient of risk aversion across the three Iowa counties under consideration. Concerning technology, they contrasted their results with a straightforward Just–Pope estimation and with the Nelson and Preckel analysis to find that each estimated similar technology structures.

The Love and Buccola approach is restrictive in the sense that CARA was imposed. Chavas and Holt (1996) developed a joint estimation method that is able to test for CARA or DARA. Applying their estimator to corn and soybean acreage allocation in the United States, and on a data set much the same as that used in their 1990 work, they assumed that the production technology was a quadratic function of allocated acres and that the utility function is $u(\pi_t, t) = \int_L^{\pi_t} \exp(\alpha_0 + \alpha_1 z + \alpha_2 z^2 + \alpha_3 t)\, dz$, where L is a lower bound on profit realizations, t is time, the α are parameters to be estimated, π_t is profit in year t, and z is a dummy variable of integration. Their analysis found strong statistical evidence for the presence of downside risk aversion and for rejecting CARA in favor of DARA.

Although the approach by Chavas and Holt does generalize the representation of risk preferences, the assumed technology was not flexible in the Just–Pope sense. Further, their specification can say little about the impact of relative risk aversion. Using Saha's (1993) expo-power utility specification, $U[\pi] = -\exp(-\beta\pi^{\alpha})$ where α and β are parameters to be estimated, Saha, Shumway and Talpaz (1994) assumed a Just–Pope technology and jointly estimated the system using maximum likelihood methods. Data were for fifteen Kansas wheat farms over the four years 1979 to 1982, and there were two aggregated input indices in the stochastic technology (a capital index and a materials index). The results supported the hypotheses of DARA and increasing relative risk aversion (IRRA). Also, the materials index was found to be risk decreasing, so risk-averse agents may have a tendency to use more fertilizer and pesticides than risk-neutral agents.

Before leaving the issue of risk estimation, a comment is warranted about subsequent use of the estimates. There may be a tendency on the part of modelers engaged in policy simulation to use without qualification risk preference structures that were identified in previous research. Newbery and Stiglitz (1981, p. 73) have shown that caution is warranted in accommodating the particular circumstances of the simulation exercise. One must ensure that the chosen risk preference structure is consistent with reasonable

levels of risk premia for the problem at hand. The set of coefficients of absolute risk aversion that give reasonable risk premia vary from problem to problem.

4.4. Econometric estimation of supply models with risk

One of the most widely agreed upon results from the theory of the firm under price uncertainty is that risk affects the optimal output level. Normally, the risk-averse producer is expected to produce less than the risk-neutral producer, ceteris paribus, and the risk-averse producer will adjust output to changing risk conditions (e.g., decrease production as risk increases). Econometric studies of agricultural supply decisions have for a long time tried to accommodate these features of the theory of the firm. There are essentially two reasons for wanting to do so: first, to find out whether the theory is relevant, i.e., to "test" whether there is response to risk in agricultural decisions; second, assuming that the theory is correct and risk aversion is important, accounting for risk response may improve the performance of econometric models for forecasting and/or policy evaluation, including welfare measurement related to risk bearing.

To pursue these two objectives, a prototypical model is to write supply decisions at time t as

$$y_t = \beta_0 + x_t'\beta_1 + \beta_2\mu_t + \beta_3\sigma_t^2 + e_t, \tag{4.3}$$

where y denotes supply, μ denotes the (subjective) conditional expectation of price, σ^2 denotes the (subjective) conditional variance of price, x represents the vector of all other variables affecting decisions, e is a random term, t indexes observations, and $(\beta_0, \beta_1, \beta_2, \beta_3)$ are parameters to be estimated (β_1 is a vector). Clearly, this formulation simplifies theory to the bone by choosing a particular functional form and, more important, by postulating that mean and variance can adequately capture the risk facing producers. Whereas more sophisticated models may be desirable, from an econometric point of view Equation (4.3) is already quite demanding. In particular, the subjective moments of the price distribution μ_t and σ_t^2 are unobserved, and thus to implement Equation (4.3) it is necessary to specify how these expectations are formed.

The specification of expectations for the first moment is a familiar problem in econometric estimation. Solutions that have been proposed range from naive expectations models (where $\mu_t = p_{t-1}$), to adaptive expectations (where μ_t is a geometrically weighted average of all past prices), to rational expectations (where μ_t is the mathematical expectation arrived at from an internally consistent model of price formation, for example). A review of price expectations formation for price levels is outside the scope of this chapter, but we note that, not surprisingly, parallel issues arise in the context of modeling variance. Behrman (1968) allowed for price risk to affect crop supply in a developing country by measuring σ_t^2 as a three-year moving average (but around the unconditional mean of price). Similar ad hoc procedures have been very common in other studies, although often with the improvement of a weighted (as opposed to simple)

average of squared deviations from the conditional (as opposed to unconditional) expectation of the price level [e.g., Lin (1977), Traill (1978), Hurt and Garcia (1982), Sengupta and Sfeir (1982), Brorsen et al. (1987), Chavas and Holt (1990, 1996)]. A more ambitious and coherent framework was proposed by Just (1974, 1976), whereby first and second moments of price are modeled to the same degree of flexibility by extending Nerlove's (1958) notion of adaptive expectations to the variance of price. This procedure has been used in other studies, including [Pope and Just (1991), Antonovitz and Green (1990), and Aradhyula and Holt (1990)]. More recently, advances have been made by modeling the time-varying variance within the autoregressive conditional heteroskedasticity (ARCH) framework of [Engle (1982)], as in [Aradhyula and Holt (1989, 1990), Holt and Moschini (1992), and Holt (1993)].

The empirical evidence suggests that risk variables are often significant in explaining agricultural production decisions. The early work by Just (1974), as well as some other studies, has suggested that the size of this supply response to risk may be quite large, but the quantitative dimension of this risk response is more difficult to assess because results are typically not reported in a standardized manner. For example, an interesting question in the context of supply response concerns the size of the likely output contraction due to risk. As model (4.3) suggests, an approximate estimate of this output reduction (in percentage terms) is simply given by the elasticity of supply with respect to the price variance σ_t^2, but this basic statistic often is not reported. As a yardstick, however, we note that for broiler production Aradhyula and Holt (1990) found a long-run price variance elasticity of -0.03, whereas for sow farrowing, the comparable long-run elasticity estimated by Holt and Moschini (1992) was -0.13.

Although such estimates may suggest a fairly sizeable production response to the presence of risk, caution is in order for several reasons. First, as is often the case in applied economic modeling, these empirical results are drawn from models that are based on individual behavior but that are estimated with aggregate data without explicit consideration of aggregation conditions. Second, insofar as producers use appropriate risk management procedures (see Section 5), the conditional variance typically used may not be measuring the relevant risk.[18] Finally, estimating response to conditional variance is inherently difficult. To illustrate this last point, consider the adaptive expectation approach that specifies the (subjective) conditional mean and the conditional variance as follows:

$$\mu_t = \sum_{k=0}^{\infty} \lambda^k (1-\lambda) p_{t-k-1}, \tag{4.4}$$

$$\sigma_t^2 = \sum_{k=0}^{\infty} \phi^k (1-\phi)[p_{t-k-1} - \mu_{t-k-1}]^2, \tag{4.5}$$

[18] For example, a producer facing price risk and using futures contracts optimally to hedge risk would be exposed only to residual basis risk, and conceivably that is what the variance terms should measure.

where usually $\lambda \in (0, 1)$ and $\phi \in (0, 1)$. These parameterizations are appealing because they make the unobservable variable a function of past realizations (which are, at least in principle, observable) in a very parsimonious way. It is known that the assumption of adaptive expectations for the mean of price is rather restrictive, and it turns out that such an assumption for the variance is even more restrictive.

By definition, if μ_t denotes the agent's conditional expectation of price, then a price-generating equation consistent with the agent's beliefs is $p_t = \mu_t + u_t$, where u_t is a random term with a zero conditional mean. Hence, an equivalent way of saying that the producer's expected price is formed adaptively as in Equation (4.4) is to say that the producer believes that price is generated by

$$p_t = p_{t-1} - \lambda u_{t-1} + u_t \tag{4.6}$$

with $\mathrm{E}[u_t \mid P_{t-1}] = 0$, where P_{t-1} denotes the entire price history up to period $t-1$. Thus, adaptive expectation for the conditional mean of price is equivalent to assuming that the agent believes that price changes follow an invertible first-order moving-average process, a rather restrictive condition.[19]

Given that Equation (4.6) is the relevant price model, the adaptive expectation model for the variance of Equation (4.5) can be rewritten as

$$\sigma_t^2 = \phi\sigma_{t-1}^2 + (1-\phi)u_{t-1}^2. \tag{4.7}$$

Note that for the model to be internally consistent the agent must believe that the random terms u_t are drawn from a distribution with mean zero and variance σ_t^2. But, as is apparent from (4.7), for most types of distributions (including the normal), σ_t^2 is bound to converge to zero as time passes. Indeed, Equation (4.7) shows that the adaptive expectation model for conditional price variance is a special case of Bollerslev's (1986) generalized ARCH (GARCH) model, specifically what Engle and Bollerslev (1986) called the "integrated" GARCH model. For this model, $\sigma_t^2 \to 0$ almost surely for most common distributions [Nelson (1990)].[20] The fact that these models imply that $\sigma_t^2 \to 0$ leads to the somewhat paradoxical situation of modeling response to risk with models that entail that risk is transitory. As Geweke (1986, p. 59) stated, "...the integrated GARCH model is not typical of anything we see in economic time series".

These undesirable modeling features are avoided if the conditional price variance is modeled by a regular GARCH model, such as the GARCH(1,1) model:

$$\sigma_t^2 = \alpha_0 + \alpha_1\sigma_{t-1}^2 + \alpha_2 u_{t-1}^2, \tag{4.8}$$

[19] See, for example [Pesaran (1987, p. 19)].

[20] Similar problems also apply to other more ad hoc parameterizations, such as that used by Chavas and Holt (1990), where $\sigma_t^2 = \sum_k \alpha_k u_{t-k}^2$ and α_k are predetermined constants satisfying $\sum_k \alpha_k = 1$.

where $\alpha_0 > 0$ bounds the conditional variance away from zero (and thus precludes $\sigma_t^2 \to 0$), and $\alpha_1 + \alpha_2 < 1$ ensures stationarity of the conditional variance process. This class of models, popular in finance studies, has been applied to agricultural supply models by Aradhyula and Holt (1989, 1990), Holt and Moschini (1992), Holt (1993), and others. Whereas this approach offers a coherent framework for modeling production response to risk, the GARCH model makes explicit the relation between conditional and unconditional variance and brings to the fore an important feature of the problem at hand. Namely, models such as (4.3) can identify response to variance only if the latter is time-varying. If, on the other hand, producers perceive variance to be relatively constant, then no response to risk can be estimated. For example, in the logic of the model (4.8), a constant variance would imply that $\alpha_1 = \alpha_2 = 0$, such that the conditional variance is the same as the unconditional variance (α_0, in such a case), and the term $\beta_3\alpha_0$ in Equation (4.3) would then be absorbed by the intercept.

We conclude this section with two observations. First, the assumption that producers perceive a constant conditional variance may not be a bad approximation. Most economic time series do seem to display ARCH properties, but the ability to forecast squared errors is usually very limited even in these models [Pagan and Schwert (1990)], and this is particularly true for the planning horizons typical of agricultural production decisions [Holt and Moschini (1992)]. Thus, in such cases conditional variance does not do much better than unconditional variance for the purpose of measuring the relevant risk; hence, identifying and estimating risk response may be too ambitious an undertaking.[21] But second, the fact that we may have trouble identifying risk response does not mean that production adjustments to risk are not present. Indeed, virtually any supply model that has been estimated without a risk term is consistent with a potentially large risk response insofar as the relevant risk is an unconditional variance that is captured by the intercept.

4.5. Risk and equilibrium in supply and production systems

The models that we have just reviewed introduce a risk variable in a single equation supply model. As mentioned earlier, representing risk in terms of a single variable (say, price variance) may be justified as an approximation to the more general EU model and will be an admissible procedure only under certain restrictive conditions (for example, normality and CARA). Whereas consideration of higher moments has been advocated [Antle and Goodger (1984)], it is arguable that such ambitions may be frustrated in most empirical applications. The single equation nature of these supply models, on the other hand, can only be a partial representation of the more complete production and supply system that may represent the agricultural producer's decision problem. Thus, generalizing risk response models to systems of equations may be desirable, and it has been

[21] A related point is that, unlike typical finance applications, agricultural supply models with risk are usually estimated with a small sample of observations.

pursued by Coyle (1992), Chavas and Holt (1990, 1996), and Saha, Shumway and Talpaz (1994), among others. Consideration of such complete supply systems is common in applied work under assumptions of certainty or risk neutrality, thanks partly to the popularization of flexible functional forms for dual representations of technology (such as profit and cost functions), which greatly simplify the derivation of coherent systems of output supply and input demand equations. Extension of this "dual" approach under risk has been explored by Coyle (1992), but because his set-up relies on a linear mean-variance objective function (which, as discussed earlier, is consistent with EU only under restrictive assumptions), it is unclear whether this dual approach is better than the corresponding "primal" approach.

The system approach typically can accommodate such integrability conditions as symmetry, homogeneity, and curvature (say, convexity in prices of the profit function). Interest in these restrictions can arise for at least two reasons. First, this set of testable restrictions may be used to validate the theoretical framework. Second, if testing the theory is not an objective, then maintaining these restrictions may be useful in improving the feasibility/efficiency of estimation, as well as improving the usefulness of empirical results for policy and welfare analysis. If one wanted to consider the integrability conditions for EU maximizing producers, what would such conditions look like? Pope (1980) pursued this question and showed that the simple symmetry and reciprocity conditions that hold under certainty need not hold under uncertainty. But, as in any optimization problem, some symmetry conditions must exist, and for the case of a producer who maximizes expected utility under price uncertainty, these conditions were characterized by Pope (1980), Chavas and Pope (1985), and Paris (1989). In general the relevant symmetry conditions will involve wealth effects (and thus will depend on risk attitudes). Restrictions on preferences, however, can reduce the symmetry and reciprocity conditions of the risk-averse case to those of the certainty case. That will happen, for example, if the utility function is of the CARA type [Pope (1980)]. Alternatively, restrictions on the technology can also reduce the symmetry and reciprocity conditions of the risk-averse case to those of the certainty case. Specifically, if the production function is homothetic, then input demands satisfy the symmetry conditions that hold under certainty; and if the production function is linearly homogeneous, then the corresponding reciprocity conditions also hold [Dalal (1990)].

A fundamental restriction of output supply and input demand functions under certainty is that of homogeneity of degree zero in prices. Thus, for example, if all input and output prices are scaled by a constant (for instance, a change of units of measurement from dollars to cents), then all real decisions are unaffected, i.e., there is no money illusion. In general the homogeneity property does not seem to hold under price uncertainty, as noted by Pope (1978) and Chavas and Pope (1985), unless restrictions are placed on preferences. Because a proportional change in all input and output prices induces a corresponding change in profit, the decisions of a producer with CARA preferences are affected by such a proportional change. On the other hand, if the producer

holds CRRA preferences, then decisions are not affected by such a proportional change in all prices.[22]

Spelling out such homogeneity conditions is quite useful, and indeed Pope (1988) used homogeneity to derive tests for the structure of risk preferences. But because homogeneity of degree zero of choice functions in prices is typically associated with the absence of money illusion, the conclusion that homogeneity need not hold under uncertainty may seem somewhat puzzling. One way to look at the problem is to recognize that the absolute risk-aversion coefficient is not unit-free; thus, for example, it is meaningless to postulate a particular numerical value for λ independent of the units of measurement of prices. If doubling of all prices were associated with halving of λ, for example, then even under CARA choices would not be affected by such a change. There is, however, a more fundamental way of looking at the homogeneity property. The crucial element here is to recognize that the vNM utility function of money, say $U(\pi)$, is best interpreted as an indirect utility function of consumer demand, such that π creates utility because it is used to purchase consumption goods. Thus, $U(\pi) \equiv V(p^c, \pi)$ where $V(p^c, \pi)$ is the agent's indirect utility function, and p^c denotes the price vector of consumption goods. In analyses of risk models, the vector p^c is subsumed in the functional $U(\cdot)$ under the presumption that these prices are held constant. Because $V(p^c, \pi)$ is homogeneous of degree zero in p^c and π, it follows that, when consumption prices are explicitly considered, the vNM utility function is homogeneous of degree zero in all prices (i.e., consumption prices, output prices, and input prices). Thus, homogeneity (i.e., lack of money illusion) must hold even under uncertainty, when this property is stated in this extended sense.

Storage opportunities introduce dynamics and require a more careful accounting for equilibrium issues as well as for expectation formation when modeling supply. In particular, because negative storage is impossible, nonlinearities are inherent in the equilibrium problem. Using U.S. soybean market data over the period 1960 to 1988, Miranda and Glauber (1993) develop an equilibrium rational expectations model that explicitly represents the behavior of producers, consumers, and storers (both private and public). They find evidence to suggest that both acres supplied and storage activities respond negatively to increased price risk. The storage result suggests that risk management institutions may facilitate efficiency by reducing impediments to intertemporal transactions.

4.6. Programming models with risk

In a number of agricultural economics applications, especially those with a normative focus, risk has been considered within suitably parameterized programming models that

[22] For example, if output and input prices are scaled by a constant $k > 0$, then profit changes from π to $k\pi$. If utility is CARA, then $-\exp(\lambda\pi) \neq -\exp(-k\lambda\pi)$, because scaling prices by k is equivalent to changing the constant coefficient of risk aversion. On the other hand, if utility is CRRA, say $U = \log(\pi)$, then scaling profit by k clearly has no effect on choices.

can readily be solved (and simulated) by appropriate computational methods. The classical quadratic programming problem of Freund (1956) maximizes a weighted linear summation of mean and variance subject to resource constraints:

$$\underset{x}{\text{Max}}\, \mu(x) - \frac{1}{2}\lambda V(x) \quad \text{such that } G(x) \leqslant 0, \tag{4.9}$$

where $\mu(x)$ and $V(x)$ are mean and variance of returns as a function of choices, $G(x) \leqslant 0$ is a vector of equality and inequality constraints, and λ measures the magnitude of risk aversion. Sharpe (1963), among others, refined the approach into a convenient and economically meaningful single-index model for portfolio choice. Applications of the method in agricultural economics include Lin, Dean and Moore (1974) and Collins and Barry (1986), both of which consider land allocation decisions. Because solving quadratic programming problems was, at one time, computationally difficult, Hazell (1971) linearized the model by replacing variance of reward with the mean of total absolute deviations (MOTAD) in the objective function. Hazell's MOTAD model has been extended in several ways by Tauer (1983), among others, and the general method has been used widely in economic analyses of agricultural and environmental issues [Teague et al. (1995)]. Risk considerations can also be introduced as a constraint, and many such programming problems go under the general rubric of safety-first optimization as studied by Pyle and Turnovsky (1970) and Bigman (1996).[23]

Given the strong relationship between time and uncertainty, risk has a natural role in dynamic optimization problems. The analytical problems associated with identifying the time path of optimal choices often requires numerical solutions for such problems. This is particularly true in agricultural and resource economics, where the necessity to accommodate such technical realities as resource carry-over may preclude stylized approaches such as the real options framework discussed previously. Stochastic dynamic programming is a discrete-time variant of optimal control methods and is robust to accommodating the technical details of the rather specific problems that arise in agricultural and natural resource economics. A standard such problem is

$$\underset{x_t}{\text{Max}} \sum_{t=0}^{T} \beta^t \mathrm{E}_0\left[\pi(x_t, y_t)\right] \quad \text{such that } y_t = f(y_{t-1}, x_{t-1}, \varepsilon_t),\ y_0 \text{ given}, \tag{4.10}$$

where T may be finite or infinite, β is the per-period discount factor, and $\pi(x_t, y_t)$ is the per-period reward. The goal is to choose, at time 0, a contingently optimal sequence, x_0 through x_T, to maximize the objective function. But the problem is not deterministic because randomness, through the sequence ε_t, enters the carry-over equation, $y_t = f(y_{t-1}, x_{t-1}, \varepsilon_t)$. This means that a re-optimization is required at each point in the time sequence. To initialize the problem, it is necessary that y_0 be known. For

[23] Note that safety-first approaches to risk modeling may be difficult to reconcile with the EU framework.

analytical convenience, Markov chain properties are usually assumed for the stochastic elements of the model. Many variants of the above problem can be constructed. For example, time could modify the per-period reward function or the carry-over function. Applications of the approach include capital investment decisions [Burt (1965)] and range stocking rate and productivity enhancement decisions [Karp and Pope (1984)].

4.7. Technology adoption, infrastructure and risk

A class of production decisions where risk is thought to play an important role is that of new technology adoption. Early work in this area, reviewed by Feder, Just and Zilberman (1985), analyzed the relationships among risk, farm size, and technology adoption. More recent studies that consider the possible impact of risk on adoption include Antle and Crissman (1990) and Pitt and Sumodiningrat (1991). The availability of irrigation has been shown to be an important risk factor for technology adoption. It both increases average productivity and reduces variability of output, and often involves community or government actions (thus emphasizing how risk management opportunities may often depend upon local institutional factors). For references to the impacts of risk and irrigation on technology adoption, with special regard to the adoption of high-yielding but flood-susceptible rice in Bangladesh, see Azam (1996), Bera and Kelley (1990), and other research cited therein. This line of research suggests that technologies are often best introduced in packages rather than as stand-alone innovations. Other work on structure includes Rosenzweig and Binswanger (1993), who studied the structural impacts of weather risk in developing countries, and Barrett (1996), who considered the effects of price risk on farm structure and productivity. In the context of hybrid maize adoption, Smale, Just and Leathers (1994) argue that it is very difficult to disentangle the importance of competing explanations for technology adoption, and suggest that previous studies may have overstated the importance of risk aversion.

The introduction of a new technology often requires a substantial capital investment, and so the functioning of credit markets plays a crucial role. For collateral-poor farmers in rural communities of the less developed world, credit is often unattainable through formal channels. For example, Udry (1994) finds that in four northern Nigeria villages more than 95 percent of borrowed funds were obtained from neighbors or relatives. One of the reasons for the importance of informal lending channels is the limited means by which formal credit providers can obtain relevant information concerning the riskiness of projects. As discussed in Ray (1998), less formal sources (such as the landlord, a local grain trader, or the village moneylender) are in a better position to judge risks and to provide credit. But, perhaps due to high default risk or to the systemic nature of risk when all borrowers are from the same village, interest rates are often very high. Bottomley (1975) developed a simple model that relates equilibrium rates to default risk. It has been suggested that moneylender market power may also affect rates but, from a survey of the literature, Ray (1998) concludes that local moneylending markets

tend to be quite competitive. However, as Bottomley (1975) pointed out, the true interest rate may often be difficult to ascertain because loans are often tied in with other business dealings such as labor, land lease, and product marketing agreements.

Faced with production and price risks, poorly performing credit markets would seem to imply inadequate investments, perhaps especially in risk-reducing technologies. On the other hand, the limited liability nature of credit may create incentives for borrowers to engage in riskier projects that are also less productive on the average, compared with the projects that would have been chosen if the credit line were not available. Basu (1992) studies the effect of limited liability and project substitution on the structure of land lease contracts.

5. Risk management for agricnltnral prodncers

The purpose of risk management is to control the possible adverse consequences of uncertainty that may arise from production decisions. Because of this inherently normative goal, stating the obvious might yet be useful: risk management activities in general do not seek to increase profits per se but rather involve shifting profits from more favorable states of nature to less favorable ones, thus increasing the expected well-being of a risk-averse individual. It should also be clear that production and risk management activities are inherently linked. Most business decisions concerning production have risk implications, and the desirability of most risk management choices can only be stated meaningfully with reference to a specific production context. As for the risk implications of production decisions, a useful classification of inputs can be made following Ehrlich and Becker (1972), who identified "self-insurance" and "self-protection" activities. Self-insurance arises when a decision alters the magnitude of a loss given that the loss occurs. Self-protection takes place when a decision alters the probability that a loss will occur. Of course, agricultural inputs may have both self-insurance and self-protection attributes; for instance, fertilizer may reduce both the probability and conditional magnitude of a crop nutrient deficiency,[24] and livestock buildings can operate in the same way upon weather-related losses. Ehrlich and Becker (1972) use this classification to show that input choices modify the demand for market insurance. Expenditures on market insurance and self-insurance substitute for each other, whereas expenditures on self-protection could actually increase the demand for market insurance.

Abstracting from self-insurance and self-protection effects of production choices, farmers usually have access to a number of other tools that have a more direct risk management role. These include contractual arrangements (e.g., forward sales, insurance contracts) as well as the possibility of diversifying their portfolio by purchasing assets

[24] In a comprehensive review of literature on crop yield variability determination, Roumasset et al. (1989) conclude that nitrogen tends to increase variability. For technology adoption, Antle and Crissman (1990) suggest that variability tends to increase initially but decrease again after more is learned about the innovation.

with payoffs correlated with the returns on production activities. Risk management decisions are obviously constrained by the given institutional and market environments, i.e., what tools and programs are actually available to the farmer. Thus, the possible incompleteness of risk markets and the imperfections of capital markets are bound to be crucial to risk management.[25] As will be discussed in this section, existing risk markets, such as contingent price markets and crop insurance, typically do not allow producers to eliminate all risk (for given production choices, it may be impossible to take market positions such that the resulting total payoff is invariant to the state of nature). Whereas this may suggest scope for welfare-increasing government intervention, it also indicates that farmers just may have to bear some residual risk.[26]

In what follows we analyze in some detail contractual relationships that a producer may enter into in order to manage price and quantity risk. In particular, we emphasize price-contingent contracts (forward, futures and options) and crop insurance contracts. Whereas the analysis hopefully will clarify the role of various risk-management tools, we should emphasize that the results of most of the models analyzed below do not translate into direct risk management recommendations. For example, given the endogeneity of many of the risks faced by producers, a discussion of risk management that takes production decisions as given is to some extent artificial, although it may be analytically useful. More generally, one should keep in mind that farmers ultimately likely care about their consumption, itself the result of an intertemporal decision. Risky production and risky prices of course imply a risky farm income, but such income uncertainty may not necessarily translate into consumption risk because borrowing/saving opportunities, as well as income from other assets and/or activities (diversification), may be used to smooth consumption over time. It is nonetheless instructive to consider certain aspects of risk management in stylized models.

5.1. Hedging with price contingent contracts

"Hedging" here refers to the acquisition of contractual positions for the purpose of insuring one's wealth against unwanted changes. As discussed earlier, output price is one of the most important sources of risk for agricultural producers. Several instruments are available to farmers of developed countries to "hedge" this price risk, notably forward contracts and price contingent contracts traded on organized futures exchanges.

[25] When capital markets are imperfect, internal funding can be very important for production decisions. For this reason, Froot, Scharfstein, and Stein (1993) argue that one of the main purposes of hedging in a business is to manage cash flow so that profitable investment opportunities that arise might be pursued. The time sequence of cash flows may also be important under the risk of business failure, as discussed by Foster and Rausser (1991).

[26] From a welfare point of view, farmers may not be the main losers from market iucompleteness. Myers (1988) showed empirically that the incompleteness may benefit producers when food demand is inelastic and may benefit consumers under other circumstances. Lapan and Moschini (1996) in a partial equilibrium framework, and Innes and Rausser (1989) and Innes (1990) in a general equilibrium framework, identified roles for second-best policy interventions when some risk markets are missing.

5.1.1. *Forward contracts and futures contracts*

The biological lags that characterize agricultural production mean that inputs have to be committed to production far in advance of harvest output being realized, at a time when output price is not known with certainty. The simplest instrument often available to farmers to deal with this price risk is a "forward contract". With such a contract a farmer and a buyer of the agricultural output agree on terms of delivery (including price) of the output in advance of its realization. For example, a farmer and a buyer can agree that a certain amount of corn will be delivered at a given time during the harvest season at the local elevator for a certain price. It is readily apparent that conditions exist under which such a contract can completely eliminate price risk. To illustrate, let $q =$ output quantity produced, $h =$ output quantity sold by means of a forward contract, $p =$ the output price at the end of the production period, $f_0 =$ the forward price quoted at the beginning of the period, and $\pi =$ the profit at the end of the period. Then the random end-of-period profit of the firm that uses forward contracts is

$$\tilde{\pi} = \tilde{p}q - C(q) + (f_0 - \tilde{p})h, \tag{5.1}$$

where $C(q)$ is a strictly convex cost function (which subsumes the effects of input prices).[27] If the farmer's utility function of profit is written as $U(\pi)$, where $U''(\cdot) < 0 < U'(\cdot)$, the first-order conditions for an optimal interior solution of an EU maximizer require

$$\mathrm{E}\big[U'(\tilde{\pi})\big(\tilde{p} - C'(q)\big)\big] = 0, \tag{5.2}$$

$$\mathrm{E}\big[U'(\tilde{\pi})(f_0 - \tilde{p})\big] = 0, \tag{5.3}$$

from which it is apparent that optimal output q^* must satisfy $C'(q^*) = f_0$. This is the "separation" result derived by Danthine (1978), Holthausen (1979), and Feder, Just and Schmitz (1980). Optimal output depends exclusively on the forward price, which is known with certainty when inputs are committed to production, and hence the production activity is riskless.

The importance of the separation result lies in the fact that the agent's beliefs about the distribution of cash and futures prices, and her degree of risk aversion, are inconsequential for the purpose of making production decisions. The agent's beliefs and her risk attitudes, however, may affect the quantity of output that is sold forward. In particular, from (5.3) it follows that

$$h^* \gtreqless q^* \quad \text{as } \mathrm{E}[\tilde{p}] \lesseqgtr f_0. \tag{5.4}$$

[27] Input prices are implicitly compounded to the end of the period using the (constant) market interest rate, so that all monetary variables in (5.1) are commensurable.

Thus, for example, a producer who believes that the forward price is biased downward (i.e., $E[\tilde{p}] > f_0$) has two ways of acting to take advantage of her information (i.e., "speculating"): she could produce more than under an unbiased forward price, while holding constant the amount sold forward; or she could decrease the amount sold forward, while holding output at the level that is optimal when the forward price is unbiased. Either action results in some uncommitted output being available at harvest time that will fetch the (risky) market price. But speculating by varying output has decreasing returns [because $C''(q) > 0$ by assumption], whereas speculating by varying the amount sold forward has constant returns. Hence, speculation here takes place exclusively by varying the amount sold forward. Similarly, changes in risk aversion, and in the riskiness of the price distribution, in this setting affect forward sales but not production decisions.

An extension of the results just discussed considers futures contracts instead of forward contracts. A futures contract is, essentially, a standardized forward contract that is traded on an organized exchange, such as the Chicago Board of Trade or the Chicago Mercantile Exchange [Williams (1986)]. A futures contract typically calls for delivery of a given quantity (say, 5,000 bushels) of a certain grade of a commodity (say, No. 2 yellow corn) at a specified delivery time (say, December of a given year) at a specified location (say, a point on the Mississippi River). Because of these features, the futures price may not be exactly suited to hedge the risk of a given producer. On the other hand, futures markets are quite liquid and hedging by using futures is readily possible for all producers, even when a local buyer offering a forward contract is not available. Using futures contracts, a producer can lock in on a price for future delivery; the problem, of course, is that this precise futures price may not be what the producer needs. Such discrepancies may be due to any one of the three main attributes of an economic good: form, time, and space.[28] Because of that, the local cash price that is relevant for the producer is not the one that is quoted on the futures market, although usually it is highly correlated with it. In addition, one should note that futures entail lumpiness (only 5,000 bu. at a time for most grains, for example) as well as transactions costs. Thus, relative to a forward contract, a futures contract is an imperfect (although possibly effective) risk-reduction instrument, i.e., the producer that uses futures contracts retains "basis risk".[29]

To illustrate hedging under basis risk, let us modify the notation of the previous section by letting f_0 = futures price quoted at the beginning of the period, $\tilde{f}$ = futures

[28] For example, the commodity grown by the producer may be of a different kind (or a different grade) than that traded on the exchange; or, the producer may realize the output at a different time than the delivery time of the contract; or, the producer may realize the output at a different location than that called for in the futures contract. Grade differences may be handled by pre-specified premiums or discounts over the futures price; differences in the type of commodity lead to the problem of "cross-hedging" of Anderson and Danthine (1981); see DiPietre and Hayenga (1982) for an application. The imperfect time hedging problem was explicitly addressed by Batlin (1983).

[29] Basis in this context refers to the difference, at the date of sale, between the (local) cash price and futures price.

price at maturity of the futures contract, and h = amount of commodity sold in the futures market. As before, $\tilde{p}$ represents the cash price at harvest time, and thus basis risk means that, typically, $\tilde{p} \neq \tilde{f}$. In general, it is difficult to fully characterize the production and hedging decisions under basis risk. Some results may be obtained, however, by restricting the relationship between cash and futures prices to be linear, as in Benninga et al. (1983):

$$\tilde{p} = \alpha + \beta \tilde{f} + \tilde{\theta}, \tag{5.5}$$

where α and β are known constants, and $\tilde{\theta}$ is a zero-mean random term that is independent of the futures price.[30] The end-of-period profit of the producer can then be represented as

$$\tilde{\pi} = (\alpha + \beta f_0 + \tilde{\theta})q - C(q) + (f_0 - \tilde{f})(h - \beta q). \tag{5.6}$$

Now, if the futures price is unbiased (i.e., if $\mathrm{E}[\tilde{f}] = f_0$), it is apparent that, for any given output q, the optimal futures hedge is $h^* = \beta q$.[31] Additional results for this basis risk case are presented in Lapan, Moschini and Hanson (1991). Because in this case random profit reduces to $\tilde{\pi} = (\alpha + \beta f_0 + \tilde{\theta})q - C(q)$, the effective (hedged) price, $\alpha + \beta f_0 + \tilde{\theta}$, is still random. Hence, under risk aversion, production takes place at a point at which marginal cost is lower than the expected price (given optimal hedging), i.e., $C'(q^*) < (\alpha + \beta f_0)$, indicating that a portion of price risk due to the basis cannot be hedged away. Because there is some residual uncertainty concerning the local cash price, the degree of risk aversion also influences optimal output. Specifically, the output level q^* is inversely related to the degree of risk aversion, as in earlier results of models of the competitive firm under price risk [Baron (1970), Sandmo (1971)]. Also, a ceteris paribus increase in nondiversifiable basis uncertainty (a mean-preserving spread of $\tilde{\theta}$) will in general decrease the optimal output level, a sufficient condition being that preferences satisfy DARA [Ishii (1977)].

It is important to realize that with basis risk, even in its special formulation of Equation (5.5), the separation result, discussed earlier for the case of forward contracts, does not apply. Because hedging does not eliminate basis risk, if the agent believes that the futures price is biased then her choice will involve the possibility of investing in two risky assets (production of output and trading in futures). Thus, if the agent believes that the futures price is biased, her optimal speculative response will entail changes in both these risky assets. For the special case of CARA preferences and of a linear basis as

[30] Actually, whereas independence is sufficient for our purposes, the slightly weaker assumption that $\tilde{f}$ is conditionally independent of $\tilde{\theta}$ is both necessary and sufficient [Lence (1995)]. Of course, for some distributions (such as the multivariate normal) these two notions of independence are equivalent. Indeed, if $(\tilde{p}, \tilde{f})$ are jointly normally distributed, then the linear basis representation in (5.5) follows.

[31] Hence, the optimal futures hedge ratio h^*/q is equal to $\beta = \mathrm{Cov}(\tilde{p}, \tilde{f})/\mathrm{Var}(\tilde{f})$, the coefficient of the theoretical regression of cash price on futures price, a result that has been used in countless empirical applications.