ESSAYS ON THE STUDY OF PRODUCTION EFFICIENCY

by

YONGSEUNG HAN

(Under the Direction of Arthur A. Snow)

ABSTRACT

In this dissertation, we present a test of the Stole-Zwiebel hypothesis that individual employees have bargaining power in negotiating compensation and that firms respond by overemploying labor and by adopting inefficient technologies. Because the hypothesis generates a production function that includes the market wage, rendering problematic the use of standard duality-based estimation procedures to test their predictions of allocative and technical inefficiencies, we develop and implement a two-step procedure to test their predictions: the first step determines whether there are allocative inefficiencies in the use of input factors, and the second determines whether any extant allocative inefficiency follows the pattern implied by the Stole-Zwiebel hypothesis.

Secondly, this dissertation proposes a way of estimating a profit system with technical and allocative inefficiencies in order to identify the source of profit inefficiency. To overcome the drawback that the existence of unobserved technical inefficiency interactive in a nonlinear form has hindered panel data estimation of the system, homogeneity in technology is introduced to separate technical inefficiency from the profit frontier. Because introducing homogeneity in technologies with multiple outputs necessitates choosing a numeraire output and a different
numeraire yields different result, a rule of thumb based on careful investigation of the data is inevitable.

Thirdly, this dissertation addresses two questions regarding the relation between allocative inefficiency, substitutability between inputs, and the choice of a quantity index. The first question is how the degree of substitutability between input factors affects the extent of allocative inefficiency and the second question is how well the degree of allocative inefficiency is captured by the Fisher and Tornqvist indices. We showed that allocative inefficiency increases with input substitutability between input factors and that both the Fisher and the Tornqvist indices increase with allocative inefficiency. However, we conclude that the Fisher index captures increased allocative inefficiency more accurately than does the Tornqvist index.

INDEX WORDS: Theory of the firm, Technical Efficiency, Allocative Efficiency, Productivity, Production, Profit, Cost, Distance Function Index, Fisher Index, Tornqvist Index
ESSAYS ON THE STUDY OF PRODUCTION EFFICIENCY

by

YONGSEUNG HAN

B.A., Yonsei University, Seoul, Korea, 1991
M.S., Texas A&M University, 2000

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008
ESSAYS ON THE STUDY OF PRODUCTION EFFICIENCY

by

YONGSEUNG HAN

Major Professor: Arthur A. Snow
Committee: Ronald S. Warren Jr.
            David B. Mustard

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. Arthur A. Snow, Dr. Ronald S. Warren Jr., and Dr. David B. Mustard because this dissertation could not be completed without assistance and support from my academic mentors. Dr. Snow, as my major professor, has showed a way of being a scholar, who is always rigorous in theoretical refinement but filled with humorous wit. I learned that a very careful use of a tilde and a hat in the proofs is just a beginning of refinement. I must confess that his insightful and clear-cut command of must-dos and must-not-dos saves my time and effort as a beginner in this research. Dr. Warren has provided me with numerous suggestions very helpful to empirical works and especially helped me conveying many findings in this research correctly to the written ones. Dr. Mustard has supported me in many ways, e.g., writing letters to support four terms of CPT, signing in the forms for tuition-waiver and an extension of I-20s as well as late registration. Moreover, he introduced me to the Association of Christian Economists and encouraged me to be in a spiritual strength.

My special thanks go to Dr. C.A. Knox Lovell and Dr. Fred Bateman. Dr. Lovell introduced me to the literature of efficiency and productivity analysis and provided me with insights into the empirical matters of business performance. I am especially indebted to him for his taking responsibility to carry me to the final stage of this dissertation even though he had retired. Dr. Bateman, in his brilliant witty mind, has always encouraged me to study hard and given me a deep understanding of my family situation.

My standing today could not be made without my father and my mother. My father, having spent many years in the Middle East and Vietnam, did his best in providing me with resources for my early education. My mother, having lived her economical life, raised me to go to college.
I also thank to my mother-in-law for her long-time prayer of my completion and to my father-in-law for his concerns. I also like to thank to my brother and his wife as well as my sister-in-law for keeping asking when I graduate.

Choonmi Choi, my beloved wife, deserves to wear regalia because she forced me to stop walking on the path of a central banker and to study economics here in the University of Georgia. Once I came for a study, she dedicated her time and effort to take care of David and Daniel, my two sons, to give me a ride, to manage financial stability, and so on. Above all, she is a company with whom I can share my joy and distress. David and Daniel are not an exception to my thanks because they have played with me.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACKNOWLEDGEMENT</td>
<td>iv</td>
</tr>
<tr>
<td></td>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2.</td>
<td>A TEST OF THE STOLE-ZWIEBEL HYPOTHESIS: AN APPLICATION TO KOREAN SAVINGS BANKS</td>
<td>6</td>
</tr>
<tr>
<td>2.1</td>
<td>INTRODUCTION</td>
<td>6</td>
</tr>
<tr>
<td>2.2</td>
<td>THE STOLE-ZWIEBEL HYPOTHESIS</td>
<td>9</td>
</tr>
<tr>
<td>2.3</td>
<td>STRATEGY FOR TESTING THE STOLE-ZWIEBEL</td>
<td>11</td>
</tr>
<tr>
<td>2.4</td>
<td>DATA AND ESTIMATION PROCEDURE</td>
<td>17</td>
</tr>
<tr>
<td>2.5</td>
<td>ESTIMATION RESULTS</td>
<td>22</td>
</tr>
<tr>
<td>2.6</td>
<td>CONCLUSIONS</td>
<td>24</td>
</tr>
<tr>
<td>3.</td>
<td>ESTIMATING A PROFIT SYSTEM WITH TECHNICAL AND ALLOCATIVE INEFFICIENCY ASSUMING HOMOGENEITY IN TECHNOLOGY</td>
<td>30</td>
</tr>
<tr>
<td>3.1</td>
<td>INTRODUCTION</td>
<td>30</td>
</tr>
<tr>
<td>3.2</td>
<td>HOMOGENEITY OF TECHNOLOGY AND SEPARABLE PROFIT FRONTIER</td>
<td>33</td>
</tr>
<tr>
<td>3.3</td>
<td>THE PROFIT SYSTEM WITH TECHNICAL AND ALLOCATIVE</td>
<td></td>
</tr>
<tr>
<td>3.4 DECOMPOSITION OF PROFIT EFFICIENCY</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>3.5 EMPIRICAL APPLICATION</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>3.6 CONCLUSION</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>4. ALLOCATIVE INEFFICIENCY, SUBSTITUTABILITY, AND THE CHOICE OF A QUANTITY INDEX</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>4.1 INTRODUCTION</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>4.2 ALLOCATIVE INEFFICIENCY AND FACTOR SUBSTITUTABILITY</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>4.3 SIMULATIONS</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>4.4 CONCLUSION</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>5. CONCLUSION</td>
<td>77</td>
<td></td>
</tr>
</tbody>
</table>

**APPENDIX**

| A. DERIVATION OF THE MEAN AND THE COVARIANCE OF $\delta$ | 87 |
| B. DERIVATION OF THE TEST STATISTIC | 89 |
| C. PROOFS OF PROPOSITIONS IN CHAPTER 2 | 90 |
| D. PROOFS OF PROPOSITIONS IN CHAPTER 3 | 92 |
| E. PROOFS OF THE LEMMAS AND PROPOSITION IN CHAPTER 4 | 94 |
CHAPTER 1

INTRODUCTION

The term efficiency may be a cliché to many ordinary people and even to some scholars and policy makers. A possible reason for a cliché may be that we hear this term too often in our daily life, e.g., energy efficiency, fuel efficiency, or even water efficiency, and so on. Even historians may have to consider whether a Southern plantation system was efficient or not when they study the demise of the plantation system.

As earliest as economics began, however, economists have known that the term efficiency is not an cliché Economists know that we all intrinsically pursue efficiency because we inevitably face the scarcity of resources. Therefore, the importance of efficiency has never been disregarded in economics.

The definition of efficiency is pretty much unilateral although the term is widely used: a comparison between observed and optimal values of something, e.g., energy, fuel, or even production process. Energy efficiency, when this definition is applied to a use of energy, is the comparison of observed and optimal values of energy. Production efficiency, when this definition is applied to production process, is the comparison between observed and optimal values of outputs and inputs that is used in the production process. However easily any definition is made, the above definition alone provides us with less fruitful lessons if we do not know how to identify the source of efficiency as well as to measure both the observed and the optimal values of outputs and inputs. For this reason, simply speaking, the study of production efficiency
is identical to the study of identifying the source of efficiency and measuring both the observed and the optimal values of output and inputs.

Identifying the source of efficiency is crucial because we cannot improve production efficiency without knowing the source of efficiency. Koopmans (1951) provided a formal definition of technical efficiency of a firm: A firm is technically efficient if and only if the firm cannot increase one output without reducing at least one other output or without increasing at least one input, or the firm cannot reduce one input without increasing at least one other input or without reducing at least one output. Debreu (1951) modeled Koopman’s definition in a way to measure radically the distance between the production frontier and the actual output values of a firm. Following Koopmans (1951) and Debreu (1951), Farrell (1957) showed how to define cost efficiency and how to decompose it into its technical and allocative components although an allocative component is constructed residually as the ratio of cost efficiency to technical efficiency.

A very sensible way of identifying allocative inefficiency was suggested by Lau and Yotopoulos (1971) in the framework of profit maximization: the shadow price approach. This approach is to restore first-order conditions of profit maximization or cost minimization by treating an allocatively inefficient firm as behaving as if it achieves its goal in accordance with the perceived prices. Due to this approach, the source of production efficiency is theoretically identified.

In identifying the source of efficiency, it should be noted that the study of production efficiency has originated from the neoclassical theory of the firm. Because it has originated from the neoclassical theory, the study of production efficiency has the following properties. As the

---

1 Separated from the studies of Koopman and Debreu, Shephard (1953) proposed a way to measure technical efficiency using inputs.
neoclassical theory of the firm provides us with a platform for the determination of the optimal quantities of output and inputs, for the size of a firm in the frameworks of profit maximization or cost minimization, and for the analytical basis for the strategic behavior of firms, so does the study of production efficiency provide whether any input or output is under- or over-utilized, whether a firm is too small or not, and whether the firm achieves profit maximization or cost minimization. Moreover, as the neoclassical theory of the firm utilizes duality property, so does the study of production efficiency.

However, identifying the source of efficiency cannot be made without measuring both the observed and the optimal values of outputs and inputs. Although the observed values of outputs and inputs are obtainable, the optimal values of output and input are not obtainable if the optimal values are not sensibly defined. The sensible way to define the optimal values is a best practice. Because the optimal values are not observed, the actual values of the best practice firm is the best, or the frontier. Then, the way of constructing the best practice firm differs in econometric approaches and non-parametric approaches. While each approach has its usefulness and weakness, the stochastic frontier approach initiated by Aigner, Lovell, and Schmidt (1977), the parametric approach first applied by Atkinson and Haverson (1980), and the data envelopment analysis developed by Charnes, Cooper, and Rhodes (1978) are commonly used.

Using any of the above approaches, the performance of any organization, either business or non-profit, is widely analyzed in economics, management science, operations research literatures. The application of these approaches to the performance of an organization ranges from the areas of post services and education to the areas of fishing and forestry, and to the environment.

---

Although the study of production efficiency has been developed over the past decades and numerous applications to measure the performance of an organization are widely made, the study of production efficiency has faced with computational difficulty in identifying the source of efficiency and measuring the efficiency. The computational difficulty is enormous and unsolved when allocative inefficiency is specified in the framework of profit efficiency. Whenever allocative inefficiency is specified with technical inefficiency in the profit maximization, we face disentanglement of both inefficiencies, rendering our effort useless. Due to this difficulty, not many studies on profit efficiency have been made. Particularly, the test of the Stole-Zwiebel hypothesis based upon the two step procedure can not be made if this difficulty is not resolved.

Although computational difficulty caused by adding allocative inefficiency to technical inefficiency can be resolved, we are not aware of how much allocative inefficiency is affected by the choice of technology, i.e., factor substitutability, or how much allocative inefficiency affects a quantity index which is frequently used to portray the productivity in the use of some quantities. That is, no study has been made to exploit the relation between allocative inefficiency, factor substitutability and the choice of a quantity index. It is unfortunate that we do not know the relation when the indices of interest are currently used in the official statistic agents to measure productivity, a change in efficiency over time. When productivity is wrongfully measured, the cost of ill-calculating index will be huge because some index is used for the basis of subsidies and future policy decisions. Here find policy implications.

Looking at the computational difficulty in the study of production efficiency and the unexploited relation between allocative inefficiency, factor substitutability, and the choice of an index, this dissertation will be devoted to provide three main chapters which will address these matters in the study of production efficiency. Chapter 2 will present one way of testing the Stole-
Zwiebel hypothesis which is alternative to the neoclassical theory of the firm. The Stole-Zwiebel hypothesis predicts that a firm responds by overemploying labor and by adopting inefficient technologies if the firm faces individual bargaining power of its employees in negotiating compensation. In chapter 2, we will develop and implement a two-step procedure to test their predictions: the first step determines whether there are allocative inefficiencies in the use of input factors, and the second determines whether any extant allocative inefficiency follows the pattern implied by the Stole-Zwiebel hypothesis. Chapter 3 will propose a way of estimating a profit system with technical and allocative inefficiencies in order to identify the source of profit inefficiency. In order to overcome the computational difficulty, we will introduce an assumption of homogeneity in technology to separate technical inefficiency from the profit frontier. Chapter 4 will address two important questions. The first question is how the degree of substitutability between input factors affects the extent of allocative inefficiency and the second question is how well the degree of allocative inefficiency is captured by the Fisher and Tornqvist indices. Chapter 5 will summarize what this dissertation has found and contributed to the study of production efficiency.
CHAPTER 2

A TEST OF THE STOLE-ZWIEBEL HYPOTHESIS:
AN APPLICATION TO KOREAN SAVINGS BANKS

2.1 INTRODUCTION

The theory of the firm has an evolutionary history in the economics literature. The neoclassical theory has addressed the determination of the optimal quantities of output and inputs as well as the size of a firm in the framework of profit maximization and cost minimization, and provides a useful analytical basis for investigating the strategic behavior of firms in product and factor markets. However, as Adam Smith pointed out in the 18th century, the assumptions that technology is given and that monitoring and coordination of factors are costless are too simple to provide a framework for answering important questions about the boundaries between firms and markets, incentive problems within firms and between firms and their investors and customers, and so on. The agency view, first formalized by Jensen and Meckling (1976), tackles incentive problems between firms and their investors, while the literature on transaction cost economics, initiated by Coase (1937), revived by Williamson (1979), and extended by Hart and Moore (1990), tackles questions about the boundaries and ownership of firms. These studies emphasize the costs of monitoring and coordination as determinants of observed institutional arrangements.

Despite this evolution of the theory of the firm, the interactive relationship among hiring, choice of technology, and organizational design had not been much attended until the work of
Stole and Zwiebel (1996a, 1996b). They emphasize the bargaining power of individual employees in negotiating compensation, and explore the subsequent effect on hiring and the choice of technology. They argue that, given employee hold-up power, a firm responds by hiring more workers to offset this power, which leads to allocative inefficiency. Technical inefficiency is introduced in this process insofar as a firm is prepared to sacrifice productive efficiency if this enhances its profitability given the wage-bargaining process. In short, the Stole and Zwiebel hypothesis predicts both technical inefficiency and allocative inefficiency originating from the intra-firm bargaining process.

Alternatively, Depken et al. (2001) hypothesize that it is the employee’s power to withhold productive effort through shirking once employed, not a worker’s power to withhold employment as in the Stole-Zwiebel hypothesis, that is the ultimate source of intra-firm inefficiencies. Allocative inefficiency arises from the institutions used to control shirking rather than from the firm hiring more workers, as in the Stole-Zwiebel model. The result of introducing profit-enhancing technical inefficiency is an increase in labor hoarding, rather than a reduction in over-employment as in the Stole-Zwiebel model. Given the conflicting hypotheses, it is interesting and important to test the Stole-Zwiebel model since the alternative hypothesis was empirically tested and shown to explain the contrasting behavior of private and collective dairy farms in Jewish Palestine.

Unfortunately, a test of the Stole-Zwiebel hypothesis is deterred by the following factors: lack of information on the bargaining environment and an inability to apply standard duality results, as Stole and Zwiebel explicitly acknowledged. First, lacking observations on the bargaining environment, researchers cannot verify a priori the assumption that wage rates are bargaining-determined rather than market-determined. Second, standard duality properties do not
hold in the Stole-Zwiebel model because the firm’s induced production function includes the market wage rate as an argument. As a consequence, any application of estimation procedures developed from standard duality theory is problematic.

We circumvent these difficulties by directly testing two key predictions of the Stole-Zwiebel model. First, their model predicts that firms over-employ labor as a means of diluting the hold-up power of employees implicit in the wage-bargaining process. Second, the distortions introduced through profit maximizing responses to intra-firm wage bargaining are characterized by a single statistic they call the front-load factor. Firms increase employment until the wage is driven down to the workers’ reservation wage where the front-load factor is equal to one.

Thus, in the first step of our procedure we test for the presence of allocative inefficiency. Using data on Korean savings banks, we find statistically significant evidence of allocative inefficiency in line with the first prediction of the Stole-Zwiebel model. Therefore, we turn to the second step and estimate each firm’s front-load factor. We find that the null hypothesis that every firm’s front-load factor is equal to one cannot be rejected. Accordingly, we conclude that the firms in our sample behave in a manner consistent with the Stole-Zwiebel hypothesis.

In the next section we review the key predictions of the Stole-Zwiebel model, and in the following section we develop our strategy for testing the model. In section IV we describe our dataset and specify our estimation procedure. Our empirical results are presented in section V. Conclusions are presented in section VI.
2.2 THE STOLE-ZWIEBEL HYPOTHESIS

Stole and Zwiebel (1996a) view the intra-firm bargaining process, absent in the neoclassical theory of the firm, as essential to understanding the firm’s behavior. Intra-firm bargaining, which determines wages and the firm’s payoff, is built on the assumption that employees and the firm split the joint surplus, revenue net of non-labor costs, relative to their respective outside options. The outside option for an employee is the reservation wage or the competitively determined market wage, \( w \), and the outside option for the firm is the outcome of a bargaining process with one less employee in the firm. In a bargaining process with \( n \) employees, equal bargaining power between the firm and its employee implies that

\[
\tilde{\pi}(n) = \tilde{\pi}(n-1) = \tilde{\pi}(0) + \int_0^{n-1} \pi(n) \, dn,
\]

where \( \tilde{\pi}(n) = F(n) - \tilde{w}(n) n \) is the firm’s profit given the output (revenue) \( F(n) \), and \( \tilde{w}(n) \) denotes an employee’s wage. Equilibrium wage and profit profiles then turn out to be

\[
\tilde{w}(n) = \frac{1}{n(n+1)} \sum_{i=0}^{n} iF_i'(i) + \frac{1}{2} w,
\]

\[
\tilde{\pi}(n) = \frac{1}{(n+1)} \sum_{i=0}^{n} \pi(n),
\]

where \( F_i'(n) \) denotes the marginal product of labor, and \( \pi(n) = F(n) - wn \) denotes neoclassical profit with \( n \) employees.

A firm, then, maximizes its payoff, given by (2b), which is the uniform average of the neoclassical profits as employment varies over \( i = 0, \ldots, n \) once the labor contract is signed and
production begins. Considering other input factors, \(x\), in the production process, the firm chooses optimal levels of labor and other inputs in accordance with the maximization problem

\[
Max_{n,x} \pi(n,x) = \frac{1}{n+1} \sum_{i=1}^{n} \pi(i,x).
\]  

(3)

Stole and Zwiebel conclude that the wage-negotiating firm acts as a neoclassical firm with the production function

\[
\tilde{F}(n,x) = \frac{1}{n+1} \sum_{i=0}^{n} F(i,x) + \frac{1}{2} w_n.
\]  

(4)

The first order conditions for the optimal levels of input factors, \((\tilde{n}^*, \tilde{x}^*)\), are given by

\[
\pi(\tilde{n}^*, \tilde{x}^*) = \tilde{\pi}(\tilde{n}^*, \tilde{x}^*)
\]  

(5a)

\[
\sum_{i=0}^{\tilde{n}^*} \frac{d\pi(i, \tilde{x}^*)}{dx} = 0,
\]  

(5b)

indicating that the wage-negotiating firm employs labor and other inputs up to the point where its profit is equal to the profit that would be earned in the bargaining-free case and the average of the marginal returns for other inputs falls to zero. Equal profits at \(\tilde{n}^*\) in (5a) imply that wages must be equal as well. That is, the firm hires labor until the bargaining-determined wage is driven down to the market-determined wage, at which point employees lose the premium gained by wage bargaining. In turn, profit maximization, which causes the wage premium to vanish, results in over-employment of labor relative to the neoclassical profit maximizing outcome, as shown in Figure 2.1. The implication of (5b) is that other inputs such as capital are under-utilized. Thus, the Stole-Zwiebel hypothesis, that wages are determined through intra-firm bargaining yields the prediction that profit maximization leads to allocative inefficiency.

The distortions introduced through the intra-firm bargaining process can be characterized by a single statistic that Stole and Zwiebel call the front-load factor:
\[ \gamma(F, n) \equiv 1 - \frac{1}{\pi(n)} \sum_{i=0}^{n} \frac{i}{n+1} \Delta \pi(i) \]  

(6a)

\[ = \frac{\tilde{x}(n)}{\pi(n)} \text{ where } 0 \leq \gamma \leq 1, \]  

(6b)

where \( \Delta \) is the first difference operator. The front-load factor measures the extent to which neoclassical profit margins are realized earlier in production process. Thus, at a given level of employment, (6b) indicates that firms prefer technologies with higher front-loading.\(^3\) It follows that a firm may choose a suboptimal technology with a higher front-load factor rather than an optimal technology with a lower front-load factor. Finally, (5a) and (6b) imply that the firm achieves its maximum profit when the front-load factor is equal to one. We shall exploit this prediction of the model in our empirical test of the Stole-Zwiebel hypothesis.

2.3 STRATEGY FOR TESTING THE STOLE-ZWIEBEL HYPOTHESIS

Since the production function (4) implied by the Stole-Zwiebel hypothesis does not yield standard duality properties, the predictions of allocative and technical inefficiencies cannot be tested using standard duality-based techniques to estimate production and cost functions. Therefore, to test the model, we develop and implement a two-step procedure: the first step determines whether there are allocative inefficiencies in the use of input factors, and the second determines whether any extant allocative inefficiency follows the pattern implied by the Stole-Zwiebel hypothesis. The first step is tantamount to finding whether neoclassical profit maximization holds. It is important to note that rejection of neoclassical profit maximization does not validate the Stole-Zwiebel hypothesis; a rejection of the neoclassical model is only a

\(^3\) In terms of Figure 1, higher front-loading would lead to a faster rise in the Stole-Zwiebel profit curve with its peak occurring an intersection higher on the neoclassical profit curve.
necessary condition for the Stole-Zwiebel model to be an explanation of the data. When this necessary condition is met, the second step of our procedure is to determine whether the front-load factor is equal to one, as predicted by the Stole-Zwiebel model.

2.3.1 A TEST FOR ALLOCATIVE INEFFICIENCY

Our test for allocative inefficiency involves a comparison of two models built on the following assumptions:

Model A  Neoclassical profit maximization holds or fails due to technical inefficiency.

Model B  Neoclassical profit maximization fails due to allocative inefficiency or due to allocative and technical inefficiency.

If Model A is the more suitable, then there is no allocative inefficiency and the wage determination process generating our data does not distort resource allocation. If Model B is the more suitable, then there is allocative inefficiency that could have its origin in mismanagement, government regulation, or wage bargaining. A log-likelihood-ratio test is performed to see which model better explains the data.

2.3.1.1 MODEL A

Model A specifies that the firm is characterized by the profit function

\[
\pi(p,w,u) \equiv \max_v \left\{ pv - w \cdot v; ye^{-u} = F(v) \right\},
\]  

(7a)
where \( y \) and \( p \) are output quantity and its per-unit price, \( v \) and \( w \) are input quantities and their prices, and \( F \) denotes the technology. Following Kumbhakar (2001), we allow for the possibility that the firm is technically inefficient by introducing the factor \( e^{-u} \). If \( u = 0 \), then \( e^{-u} = 1 \) and \( y = F(v) \), and the firm is technically efficient; if \( u < 0 \), then \( e^{-u} > 1 \) and \( y < F(v) \), and the firm is technically inefficient, operating in the interior of its technology set. With this convention, the profit function can be redefined as

\[
\pi(p^e, w) \equiv \max_v \left\{ \left( p^e \right)^{y} - w \cdot v; \ y = F(v) \right\}.
\] (7b)

This is the firm’s actual profit function, which we estimate with our data by adopting the translog form

\[
\ln \pi(p^e, w) = \alpha_0 + \sum_{k=1}^{K} \alpha_k \ln w_k + \alpha_y \ln \left( p^e \right) + \frac{1}{2} \sum_{k=1}^{K} \sum_{j=1}^{K} \alpha_{kj} \ln w_k \ln w_j + \alpha_{yy} \ln \left( p^e \right) \ln \left( p^e \right) + \sum_{k=1}^{K} \alpha_{ky} \ln w_k \ln \left( p^e \right),
\] (8a)

with the symmetry restrictions \( \alpha_{kj} = \alpha_{jk} \) and \( \alpha_{ky} = \alpha_{yk} \) imposed. By applying the Shepard-Uzawa-McFadden lemma to (8a), we obtain the cost-share equations

\[
S_k = -\alpha_k - \sum_{j=1}^{K} \alpha_{kj} \ln w_j - \alpha_{ky} \ln \left( p^e \right) \ \forall k
\] (8b)

To avoid singularity, we omit the revenue-share equation and estimate the system of equations (8a) and (8b). Since the profit function is homogenous of degree one in input prices and the technical-inefficiency-adjusted output price \( (p^e) \), it is necessary to impose the parameter restrictions implied by linear homogeneity,

\[
\sum_{k=1}^{K} \alpha_k + \alpha_y = 1, \sum_{k=1}^{K} \alpha_{ky} + \alpha_{yy} = 0, \text{ and } \sum_{j \neq k}^{K} \alpha_{kj} + \alpha_{ky} = 0 \ \forall k.
\] (9)
2.3.1.2 MODEL B

Model B adds the possibility that the firm fails to achieve allocative efficiency. To implement the standard duality-based estimation procedure, we adopt the shadow price approach developed by Lau and Yotopoulos (1971), used by Atkinson and Halvorsen (1980), and extended by Atkinson and Cornwell (1994). This approach ensures that the first-order conditions for profit maximization are satisfied by introducing shadow input prices \( w^s = \theta w \), where \( \theta \) is a vector that distorts the market prices \( w \) and leads to the first-order conditions

\[
\frac{\partial F}{\partial v_k} \forall k.
\]

(10)

If \( \theta = 1 \), then the shadow prices for inputs are the same as their market prices, and there are no allocative inefficiencies. However, if \( \theta \neq 1 \), then the shadow price vector differs from the market price vector, and the firm employs an allocatively inefficient input mix.

The shadow profit function is defined by

\[
\pi^s (pe^u, w^s) \equiv \max_v \left\{ (pe^u) y - w^s \cdot v; \ y = F(v) \right\},
\]

(11)

and is related to actual profit, \( \pi^u \), in the following way:

\[
\pi^u = (pe^u) F(v) - \sum_{k=1}^{K} w_k v_k
\]

\[
= \pi^s (pe^u, w^s) - \sum_{k=1}^{K} (w_k - w_k^s) v_k
\]

(12)

\[
= \pi^s (pe^u, w^s) \left[ 1 - \sum_{k=1}^{K} \frac{(1-\theta_k)}{\theta_k} S^s_k \right],
\]
where \( S_k^s = \frac{\partial \ln \pi^s(pe^s, w^s)}{\partial \ln w_k^s} \) is the shadow cost share for the kth input. The actual cost share for the kth input is \( \frac{w_k V_k}{\pi^a} \).

To estimate (12), we adopt the translog form for the shadow profit function. Thus, we estimate

\[
\ln \pi^a = \ln \pi^s(pe^s, w^s) + \ln \left[ 1 - \sum_{k=1}^{K} \left( \frac{1 - \theta_k}{\theta_k} S_k^s \right) \right], \quad (13a)
\]

where \( \ln \pi^s(pe^s, w^s) \) and \( S_k^s \) are given by (8a) and (8b), respectively, with \( w^s \) replacing \( w \), together with the actual cost-share equations

\[
S_k^a \equiv \frac{w_k V_k}{\pi^a} = \left[ 1 - \sum_{k=1}^{K} \left( \frac{1 - \theta_k}{\theta_k} S_k^s \right) \right]^{-1} \left( -S_k^s \frac{1}{\theta_k} \right) \quad \forall k. \quad (13b)
\]

2.3.2. A TEST OF THE SIZE OF THE FRONT-LOAD FACTOR

Our empirical implementation of Models A and B, reported on below, reveals statistically significant evidence of allocative inefficiency and shows that Model B better explains our data. Hence, we proceed to estimate the front-load factor for the firms in our sample. As the front-load factor is the ratio of Stole-Zweibel profit to neoclassical profit, and the former is itself based on a sum of neoclassical profits, we use Model B to derive predicted neoclassical profit for alternative levels of employment.

Construction of the Stole-Zwiebel profit for a firm with \( n \) employees involves summing the neoclassical profit associated with \( i \) employees for \( i=0, 1, \ldots, n \). Since our data set does not have
every possible number of employees represented, we estimate the relation between the predicted neoclassical profit and the observed employment for the firms in our sample as

\[ \ln \pi_j = \beta_1 \text{emp}_j + \beta_2 \text{emp}_j^2 + \beta_3 \text{TE}_j, \quad (14) \]

where \( \pi_j \) denotes the predicted neoclassical profit of firm \( j \), \( \text{emp}_j \) denotes the number of employees at firm \( j \), and \( \text{TE}_j = e^{-x_j} \) denotes the firm’s technical inefficiency, which can be viewed as a proxy for a firm’s choice of technology. Once this empirical relation is estimated, we can calculate the neoclassical profit for a firm with any number \( i \) of employees, and then using equation (2b) we can calculate the Stole-Zweibel profit for a firm with \( n \) employees. We can then calculate the firm’s front-load factor using equation (6b). Then, using the equivalence of \( \gamma(F, n) = 1 \) to \( \pi(n) = \tilde{\pi}(n) \), the null hypothesis for an individual firm \( j \) with \( n \) employees is

\[ H_0: \delta_j \equiv \ln \gamma_j \equiv \ln \tilde{\pi}_j(n) - \ln \pi_j(n) = 0. \quad (15) \]

A distributional assumption about the estimated front-load factor is crucial to this test because, without such an assumption, the estimated covariance or correlation could not provide a measure of how close the two series, \( \ln \pi(n) \) and \( \ln \tilde{\pi}(n) \), are. Accordingly, we assume that the estimated front-load factor follows a normal distribution. This assumption is justified given the usual assumption that the error term appended to equation (14) is normally distributed. Then, \( \ln \pi_j(n) \) follows the standard normal distribution, and so does \( \ln \tilde{\pi}_j(n) \) because the sum of normally distributed random variables is normally distributed. Then, the distribution of the vector \( \hat{\delta} \) of estimated values for \( \delta_j \) for each firm follows a normal distribution with mean \( \mu = E(\hat{\delta}) \) and the covariance \( \Sigma = \text{Cov}(\hat{\delta}) \). Derivations of the mean and the covariance are provided in Appendix 1.
With these distributional assumptions, our test of the hypothesis that the front-load factor equals one is performed at the industry level:

\[ H_0 : \delta_i = 0, \cdots, \delta_j = 0, \cdots, \delta_N = 0 \]  \hspace{1cm} (16)

for the N firms in our sample. This is a strong test in that all the firms’ front-load factors are hypothesized to be unitary. Hypothesis (15) concerning an individual firm’s front-load factor could, in principle, be tested if the joint hypothesis (16) is rejected. However, there is insufficient data in our sample to carry out these tests, and we are limited to testing the joint hypothesis (16). The test statistic for this hypothesis follows the F distribution,

\[ TS_i = \delta^T \hat{\Sigma}^{-1} \delta \]  \hspace{1cm} (17)

where \( \delta^T = [\delta_1, \cdots, \delta_N] \). Derivation of the test statistic is provided in Appendix 2.

2.4 DATA AND ESTIMATION PROCEDURE

2.4.1 DATA

We implement the tests developed in the previous section using data on Korean savings banks. The banking industry provides two unique characteristics that facilitate a test of the Stole-Zwiebel hypothesis. First, the recent banking industry, at least in Korea and in the U.S., is marked by competition with regulatory supervision. Competitiveness seems to drive banks to be more efficient in their use of inputs\(^4\). Regulatory supervision provides rich datasets with information on banks’ profits and portfolios. Second, the banking industry is a financial intermediary which transforms various financial and physical resources into loans and

\[^4\text{Kumbhakar and Tsionas (2005) find that banks, on average, do not seem to have significant relative price distortions. For details, see p. 377.}\]
investments. As Sealey and Lindley (1977) pointed out, failure to consider the financial intermediary function of banks has led researchers to misidentify outputs and inputs and to analyze the technical aspects of production and cost incorrectly. This observation has led to the development of the financial intermediation approach that is now most commonly followed in this literature.

In this approach, variables are defined in the following way. First, the input variables are labor \( (v_1) \), deposits \( (v_2) \) and purchased funds \( (v_3) \). The corresponding prices of the inputs are constructed as follows: the price of labor \( (w_1) \) is the sum of salaries and employment benefits divided by the number of employees; the price of deposits \( (w_2) \) is the interest paid on deposits divided by the volume of deposits; and the price of purchased funds \( (w_3) \) is the sum of interest payments on borrowings and bonds divided by the amount of purchased funds. Second, we follow Berger and Mester (2003) and introduce the quasi-fixed inputs they suggest to control for special characteristics of the banking industry: off-balance sheet items \( (z_1) \) to control for the quality of loans, on the assumption that credit risks increase as the size of loans increase\(^5\); financial capital \( (z_2) \) to control for regulatory supervision, on the assumption that banks must meet regulatory capital requirements; and physical capital \( (z_3) \) to control for the difficulty of measuring its price. Third, the output variable is defined to be bank revenues \( (y) \), and its price \( (p) \) is operating revenue divided by total assets net of physical capital. Lastly, variable profit \( (\pi^a) \) is defined as operating revenue net of variable costs. Accounting for the quasi-fixed inputs, the short-run variable-profit function is

---

\(^5\) To control for the quality of loans, other studies have used different variables. Hughes and Mester (1993) and Mester (1996) used non-performing loans, while Berg et al. (1992) used loan losses.
\[ \pi(p \in, w, z) \equiv \max_v \left\{ \left( p \in^v \right) y - w \cdot v; y = F(v, z) \right\}, \quad (18) \]

and the empirical profit and share equations (8a)–(8b) and (13a)–(13b) are also modified to include the quasi-fixed inputs.

The data are taken from information collected by the Korean Financial Supervisory Service (FSS). It is required that all savings banks submit annual reports to the FSS. We use data for the years 2002-2005 and, as we are estimating a translog profit function, our sample is limited to the 92 banks reporting positive profit. All data are deflated by the GDP deflator, as is usual in this literature [e.g. Berger and Mester (1997) and Wheelock and Wilson (1999)]. Table 2.1 provides the descriptive statistics for our sample.

2.4.2 ESTIMATION PROCEDURE

When panel data are available, fixed-effect panel estimation has been commonly used to identify an individual, firm-specific unobserved effect as technical inefficiency since Schmidt and Sickles (1984)\(^6\). Along with fixed-effect estimation, allocative inefficiency can also be parameterized as Atkinson and Cornwell (1994) suggested in their estimation of a cost system.

However, in estimating the profit system (13), it is extremely difficult computationally to separate an individual firm-specific unobserved effect from the other explanatory variables, because the technical inefficiency variable, \( u \), is embedded in the profit function. This computational difficulty explains why the profit system with technical and allocative inefficiencies has not been successfully estimated with panel data.

---

\(^6\) If only cross-section data are available, the stochastic frontier analysis is most commonly used since Aigner, Lovell and Schmidt (1977).
Here, we suggest that one way to avoid this computational difficulty is to impose the assumption that the production process is homogeneous of degree \( r < 1 \) in the variable input factors. We exploit this assumption in our estimation procedure by taking advantage of the following propositions\(^7\).

**Definition**

The normalized variable profit function, \( \hat{\pi}(w, z) \), is defined as

\[
\hat{\pi}(w, z) \equiv \max_v \{ F(v, z) - \sum_{k=1}^K w_k v_k \}.
\]

**Proposition 1**

\( \hat{\pi}(tw, z) = t^{r-1} \hat{\pi}(w, z) \) if and only if \( F(v, z) \) is homogeneous of degree \( r \) in \( v \).

**Proposition 2**

\( \pi(pe^u, w, z) = (pe^u)^{\frac{1}{1-r}} \hat{\pi}(w, z) \) if and only if \( F(v, z) \) is homogeneous of degree \( r \) in \( v \).

**Proposition 3**

Linear homogeneity of \( \pi(pe^u, w, z) \) in \( (pe^u, w) \) is equivalent to \( \hat{\pi}(w, z) \) being homogeneous degree \( \frac{r}{r-1} \) in \( w \) if and only if \( F(v, z) \) is homogeneous of degree \( r \) in \( v \).

When the production technology is homogenous of degree \( r \) in the variable input factors, Proposition 1 identifies the homogeneous structure of the normalized profit function, Proposition 2 shows that the technical-efficiency-adjusted output price can be separated from the profit function as a factor multiplying the normalized profit function, and Proposition 3 establishes that the linear homogeneity property of the profit function is equivalent to homogeneity of degree

\(^7\) Propositions 1 and 2 were established by Lau (1978) in Theorems II-1 and III-1, respectively. For completeness, we provide brief proofs of these propositions in Appendix C.
\( r/(r-1) \) for the normalized profit function. Using these results, equation (13a), with the quasi-fixed inputs included, can be respecified as

\[
\ln \pi'' = \alpha_0 \ln p + \ln \hat{\pi}^S (w^S, z) + \ln \left[ 1 - \sum_{k=1}^{K} \frac{(1-\theta_k)}{\theta_k} S_k^S \right] + u, \tag{19}
\]

where \( \alpha_0 = 1/(1-r) \), \( S_k^S = -\alpha_k - \sum_{j=1}^{3} \alpha_{kj} \ln w_j^S \), and

\[
\ln \hat{\pi}^S (w^S, z) = \sum_{k=1}^{3} \alpha_k \ln w_k^S + \sum_{m=1}^{3} \beta_m \ln z_m + \frac{1}{2} \left[ \sum_{k=1}^{3} \sum_{j=1}^{3} \alpha_{kj} \ln w_k^S \ln w_j^S + \sum_{m=1}^{3} \sum_{l=1}^{3} \beta_{ml} \ln z_m \ln z_l \right] + \sum_{k=1}^{3} \sum_{m=1}^{3} \lambda_{km} \ln w_k^S \ln z_m
\]

with the symmetry restrictions, \( \alpha_{kj} = \alpha_{jk} \) and \( \beta_{ml} = \beta_{lm} \). The cost-share equations are given by (13b) with \( K=3 \) and \( S_k^S = -\alpha_k - \sum_{j=1}^{3} \alpha_{kj} \ln w_j^S \). Finally, Proposition 3 indicates that the normalized profit function is homogeneous of degree \( (1-\alpha_o) \), implying the parameter restrictions

\[
\alpha_0 + \sum_{k=1}^{3} \alpha_k = 1, \sum_{j=1}^{3} \alpha_{kj} = 0 \text{ for } k = 1, 2, 3, \tag{20}
\]

\[
\sum_{k=1}^{3} \lambda_{km} = 0 \text{ for } m = 1, 2, 3, \text{ and } \sum_{k=1}^{3} \sum_{j=1}^{3} \alpha_{kj} = 0.
\]

The system of equations listed in (8a) and (8b) with quasi-fixed inputs included is given by (19) and (13b) with \( \theta_k \) set equal to one for all \( k \) to yield the profit system with only technical inefficiency.\(^8\)

We estimate the equation systems by iterative nonlinear least squares in TSP, invoking the heteroskedastic-consistent standard errors option after the classical, additive, normal error terms

---

\(^8\) Given the short duration of our panel, technical inefficiency is assumed to be time-invariant.
are appended to the equations. The parameterized allocative inefficiency terms are not estimated as firm-specific, since our dataset is subject to degrees-of-freedom limitations.

2.5 EMPIRICAL RESULTS

The results of estimating the system of equations in (19) and (13b) for Model A \( \theta_k = 1 \ \forall k \) and for Model B (\( \theta_k \) unrestricted) are reported in Table 2.2. It is expected that allocative inefficiencies affect the estimates of the coefficients, particularly the estimates of the coefficients on the input prices, because those allocative inefficiencies are parameterized. And it is also expected that the allocative inefficiencies are less than 1 because the banking industry is competitive enough to reduce allocative inefficiency. These a priori expectations are satisfied: parameter estimates are quite different in the two models and allocative inefficiencies are small. The small magnitude of allocative inefficiency is consistent with the findings of Kumbhakar and Tsionas (2005). Note that the coefficient of output price is greater than one in both models. This is consistent with the assumption that the technology is homogeneous of degree \( r \), since with \( r < 1 \) (that is, with production subject to decreasing returns to scale in the variable factors) the theoretical value for \( \alpha_0 \) is greater than one.

With these estimates, a log-likelihood test of the null hypothesis

\[ H_0 : \theta_1 = \theta_2 = \theta_3 = 1 \]  

that all of the coefficients of allocative inefficiency are unitary is performed. The test of this null hypothesis allows us to determine which model better explains the data, the unrestricted Model B or the restricted Model A. The test statistic, which follows a \( \chi^2 \) distribution with degrees of freedom equal to the number of restrictions, 3, is
\[ TS_2 = 2(\text{LogL}_B - \text{LogL}_A) = 208.424 \]  

(22)

where \( \text{LogL}_A \) is the maximum value of the likelihood function for Model A and \( \text{LogL}_B \) is the maximum value of the likelihood function for Model B. The \( \chi^2 \) critical value at the 0.01 significance level is 11.34. Thus, the null hypothesis is strongly rejected, and we conclude that Korean banks are subject to allocative inefficiency to some degree.\(^9\) Moreover, the estimated shadow wage rate is 53 percent of the actual wage rate, consistent with the prediction of overemployment.

Given allocative inefficiency, fixed-effect panel data estimation is applied to (14) in order to estimate the empirical relation between neoclassical profit and the level of employment.\(^10\) The regression result is

\[
\ln \pi_j = 0.008 \text{emp}_j - 0.00001 \text{emp}_j^2 + 53.608 \text{TE}_j
\]

\[
(3.605) \quad (-2.116) \quad (50.651)
\]

where t-values are given in parenthesis, and adjusted \( R^2 = 0.923 \). These results identify a quadratic relation between neoclassical profit and the level of employment, controlling for the choice of technology. This inverse U-shape relation, although moderate, is consistent with the Stole-Zweibel model. Using the empirical relation between neoclassical profit and the level of employment, each firm’s front-load factor is constructed using the current level of employment and the industry level of average technical efficiency. The results are shown in Figure 2.2.

Given these estimated front-load factors, we test the null hypothesis (16) that all of these factors are equal to one. The test statistic, which follows an F distribution with 364 and 273

\(^9\) An industry level of technical inefficiency is measured to be 17.4% in Model A and 14.6% in Model B, respectively. This result is close to the mean of 16% for U.S. banks surveyed by Berger and Humphrey (1997).

\(^10\) The Hausman test is performed to choose fixed-effect estimation rather than random-effect estimation.
degrees of freedom, is 0.481. The F critical value at the 0.05 level is 1.207. Thus, the null hypothesis that all the front-load factors are unitary cannot be rejected.

2.6 CONCLUSIONS

We develop and implement a two-step procedure to test the Stole-Zwiebel hypothesis that intra-firm wage bargaining explains observed allocative and technical inefficiencies in the behavior of firms. Devising a test is problematic since their model does not admit standard, duality-based estimation procedures. We overcome this difficulty by directly testing the two key predictions of their model. First, they argue that firms respond to the employee hold-up power implicit in wage bargaining by over-employing labor, leading to allocative inefficiency. Technical inefficiency is a potential by-product as a firm is willing to sacrifice productive efficiency if that will enhance its profitability given the wage-bargaining process. Second, these distortions can be characterized by a single statistic, the front-load factor, which is based on the neoclassical profit that would be earned at each level of employment up to the actual level. Firms expand employment until the bargaining-determined wage is driven down to the market-determined wage where the front-load factor is equal to one.

Thus, in the first step of our procedure we test for the presence of allocative inefficiency. If there is no evidence of allocative inefficiency, then we can reject the Stole-Zwiebel hypothesis. By adopting the maintained hypothesis that the production technology is homogeneous in the variable input factors, we implement this step using data on Korean savings banks to estimate a translog profit function that allows for both allocative and technical inefficiency. Our results show statistically significant evidence of allocative inefficiency. Therefore, we proceed to the
second step and derive an estimate of each firm’s front-load factor by estimating the relation between the predicted neoclassical profit and the observed employment for the firms in our sample. Our results from this step show that the null hypothesis that every firm’s front-load factor is equal to one cannot be rejected. On the basis of this evidence, we conclude that the employment behavior of Korean savings banks is consistent with the key predictions of the Stole-Zwiebel model of intra-firm wage bargaining.
Table 2.1  Mean and standard deviations of in-sample Korean savings banks

<table>
<thead>
<tr>
<th></th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price of labor ($w_1$)</td>
<td>0.028</td>
<td>0.031</td>
<td>0.034</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.010)</td>
<td>(0.016)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Price of deposits ($w_2$)</td>
<td>0.065</td>
<td>0.057</td>
<td>0.051</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.005)</td>
</tr>
<tr>
<td>Price of purchased funds ($w_3$)</td>
<td>0.012</td>
<td>0.010</td>
<td>0.010</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>(0.022)</td>
<td>(0.019)</td>
<td>(0.019)</td>
<td>(0.036)</td>
</tr>
<tr>
<td>Price of output ($p$)</td>
<td>0.118</td>
<td>0.123</td>
<td>0.103</td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td>(0.030)</td>
<td>(0.040)</td>
<td>(0.023)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Off-BS ($z_1$)</td>
<td>6,516</td>
<td>12,079</td>
<td>9,061</td>
<td>9,599</td>
</tr>
<tr>
<td></td>
<td>(10,242)</td>
<td>(14,094)</td>
<td>(15,001)</td>
<td>(18,161)</td>
</tr>
<tr>
<td>Financial capital ($z_1$)</td>
<td>16,419</td>
<td>14,762</td>
<td>15,154</td>
<td>15,399</td>
</tr>
<tr>
<td></td>
<td>(17,649)</td>
<td>(16,794)</td>
<td>(18,480)</td>
<td>(17,179)</td>
</tr>
<tr>
<td>Physical capital ($z_3$)</td>
<td>11,774</td>
<td>12,992</td>
<td>15,297</td>
<td>18,471</td>
</tr>
<tr>
<td></td>
<td>(11,779)</td>
<td>(12,180)</td>
<td>(15,424)</td>
<td>(21,093)</td>
</tr>
<tr>
<td>Variable profit ($\pi$)</td>
<td>9,148</td>
<td>13,515</td>
<td>13,028</td>
<td>15,469</td>
</tr>
<tr>
<td></td>
<td>(14,651)</td>
<td>(17,862)</td>
<td>(16,994)</td>
<td>(20,883)</td>
</tr>
</tbody>
</table>

Numbers in parenthesis are standard deviations. Prices are measured in percentage and others are measured in million won.
<table>
<thead>
<tr>
<th>Variables (parameters)</th>
<th>Model A</th>
<th>Model B</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln p</td>
<td>2.2341 (0.0869)</td>
<td>1.2783 (0.0789)</td>
</tr>
<tr>
<td>ln w_t</td>
<td>0.0744 (0.0521)</td>
<td>0.1184 (0.0329)</td>
</tr>
<tr>
<td>ln w_t</td>
<td>-1.2054 (0.0649)</td>
<td>-0.3700 (0.0730)</td>
</tr>
<tr>
<td>ln z_t</td>
<td>-1.0131 (0.0313)</td>
<td>-0.0401 (0.0250)</td>
</tr>
<tr>
<td>ln z_t</td>
<td>-0.1031 (0.0299)</td>
<td>-0.2670 (0.0307)</td>
</tr>
<tr>
<td>ln w_t ln w_t</td>
<td>-0.0241 (0.0161)</td>
<td>0.0098 (0.0055)</td>
</tr>
<tr>
<td>ln w_t ln w_t</td>
<td>0.0926 (0.0182)</td>
<td>0.0291 (0.0084)</td>
</tr>
<tr>
<td>ln w_t ln w_t</td>
<td>0.0146 (0.003)</td>
<td>0.0018 (0.001)</td>
</tr>
<tr>
<td>ln w_t ln w_t</td>
<td>-0.1905 (0.0285)</td>
<td>-0.0765 (0.018)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>-0.0335 (0.0039)</td>
<td>-0.0038 (0.0038)</td>
</tr>
<tr>
<td>ln z_t ln z_t</td>
<td>0.0088 (0.0049)</td>
<td>0.0140 (0.0042)</td>
</tr>
<tr>
<td>ln z_t ln z_t</td>
<td>-0.0037 (0.0032)</td>
<td>-0.0031 (0.0027)</td>
</tr>
<tr>
<td>ln z_t ln z_t</td>
<td>-0.0004 (0.0014)</td>
<td>0.0010 (0.0013)</td>
</tr>
<tr>
<td>ln z_t ln z_t</td>
<td>0.0092 (0.0234)</td>
<td>0.0295 (0.0202)</td>
</tr>
<tr>
<td>ln z_t ln z_t</td>
<td>-0.0076 (0.0101)</td>
<td>-0.0044 (0.0086)</td>
</tr>
<tr>
<td>ln z_t ln z_t</td>
<td>0.0825 (0.0084)</td>
<td>0.0900 (0.0072)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>0.0028 (0.0086)</td>
<td>-0.0017 (0.0074)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>0.0196 (0.0221)</td>
<td>0.0134 (0.0195)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>-0.0200 (0.0209)</td>
<td>-0.0202 (0.0178)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>-0.0027 (0.0087)</td>
<td>0.0002 (0.0075)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>-0.0021 (0.0215)</td>
<td>-0.0149 (0.0193)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>0.0151 (0.02)</td>
<td>0.0147 (0.0172)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>-0.0001 (0.0015)</td>
<td>0.0014 (0.0013)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>-0.0175 (0.005)</td>
<td>0.0016 (0.0041)</td>
</tr>
<tr>
<td>ln w_t ln z_t</td>
<td>0.0049 (0.0032)</td>
<td>0.0054 (0.0028)</td>
</tr>
<tr>
<td>(θ_1)</td>
<td>0.5293 (0.2538)</td>
<td></td>
</tr>
<tr>
<td>(θ_2)</td>
<td>0.2378 (0.0687)</td>
<td></td>
</tr>
<tr>
<td>(θ_3)</td>
<td>0.2212 (0.2076)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2 Parameter estimates of Model A and Model B

Standard Errors in parenthesis
Figure 2.1 Profit maximizing choice of labor

![Graph showing profit maximizing choice of labor]

- Profit (Neoclassical)
- Profit (S-Z)

Profit vs. Labor with points $n^*$ and $\tilde{n}^*$.
Figure 2.2  A firm’s constructed front-load factor
CHAPTER 3

ESTIMATING A PROFIT SYSTEM WITH TECHNICAL AND ALLOCATIVE INEFFICIENCY ASSUMING HOMOGENEITY IN TECHNOLOGY

3.1 INTRODUCTION

Profit efficiency is an overall indicator of a firm’s performance which is greatly influenced by all the exogenous factors surrounding a firm, as well as its efficiency in using its production process. By “overall” we mean that profit efficiency comprises both the revenue (output) and cost (input) aspects of a firm’s decisions. Thus, by examining the profit efficiency of a firm one avoids the ironic possibility of concluding that “Surgery was successful, but the patient died.” That is, one avoids the mistake of finding that a firm, which soon becomes bankrupt because of poor revenue inefficiency, is efficient when cost efficiency alone is analyzed. The mistake is not purely imaginary, as studies of the U.S. banking industry (Berger et al. 1993, Berger and Mester 1997) have found that profit efficiency is only slightly correlated with cost efficiency.

Previous studies of profit efficiency have mainly focused on the effect of changes in various exogenous factors on a firm’s profit. Estimation of a single equation, either by a stochastic frontier approach or by other approaches such as a data envelopment analysis (DEA), centers on the analysis of this effect. That is, exogenous factors are estimated either jointly or separately
with the technical inefficiency term ($u$) to evaluate the effect of exogenous factors on the performance of firms.

Typically, a two-step estimation procedure is employed that regresses profit efficiency scores obtained in the first-step on a set of variables characterizing exogenous effects. For the banking industry, Berger and Mester (1997) synthesized previous studies and classified exogenous effects into 6 categories (size, organizational form and governance, other bank characteristics, market characteristics, geographical restrictions on competition, and regulation). Subsequent studies contain variants of this approach with special focus on the effect of organizational control and distance on the efficiency of the affiliates of U.S. banks (Berger and DeYoung, 2001), on ownership of banks in transition countries (Bonin et al. 2005), and on smaller banks in the U.S. (Akhigbe and McNulty, 2005). The two-step procedure, however, is problematic because estimates of both the technology parameters and inefficiency in the first stage estimation are likely to be biased. That is, exogenous factors might be endogenous (Berger and Mester, 1997, p 911). More importantly, the two-step procedure leads to a potentially serious bias of $u$, unless a scaling property is considered (Wang and Schmidt, 2002).

An alternative approach is to estimate the exogenous factors jointly with the inefficiency term directly in a single stage (Berger and Mester, 2003). This estimation method is different from the one-step procedure in which the inefficiency term is a function of exogenous factors in the composite error components. The assumption that the exogenous factors are linear in the functional form has a weak theoretical foundation in terms of the specification of a functional form.

---

11 In the two-step procedure, an efficiency score as a dependent variable can be obtained by DEA. An early example in the banking literature is Aly et al. (1990), although it is in the context of cost minimization.
12 Neither Tobit regression in the second stage nor MLE-based truncated normal distribution in the first stage corrects this bias unless a scaling factor (or a correcting factor) is introduced. For details, see Simar and Wilson (2008, pp. 91-95).
While the effect of exogenous factors on a firm’s efficiency has been widely studied, identifying the source of profit inefficiency has not been fully successful because of the formidable difficulty in estimation. The study by Huang (2000) on Taiwanese banks is subject to the criticism that technical inefficiency was inappropriately removed from the normalized shadow profit function. Berger et al. (1993) and DeYoung and Nolle (1996), studying U.S. banks and foreign banks in the U.S. and utilizing the Fuss normalized quadratic profit function, do not separate technical and allocative inefficiencies.

The error component approach, in which any inefficiency is characterized as an error term, entangles technical and allocative inefficiency in the profit function. Simply put, if the sources of technical and allocative inefficiency are characterized as error terms \( u \) and \( A \), respectively, a profit function has three-way error components, i.e., \( u + A + v \) where \( v \) is a random error. Kumbhakar (2001), using the shadow price approach to allocative inefficiency rather than the error component approach, proposed the estimation of share equations only (that is, dropping the profit function which has error components). However, an extension of his proposal to panel data has not yet been implemented, and one possible price of dropping the profit function is a loss in the precision of the resulting estimates.

Difficulty in empirical estimation remains when panel data are available. A flexible functional form resulting in a quadratic form for technical inefficiency in the profit function fails to incorporate technical inefficiency \( u \) as a unobserved firm-specific effect. Dropping the profit function is not sufficient to estimate the profit system, as Atkinson and Cornwell (1998) noted.

In this paper, I propose a way to estimate a profit system with technical and allocative inefficiencies by assuming homogeneity in technology and estimate the system with data on Koran savings banks. In a nutshell, assuming homogeneity in technology enables me to separate
the profit frontier and technical inefficiency with a fixed-effects panel data estimation technique. However, it should be emphasized that the homogeneity assumption on technology with multiple outputs inevitably necessitates choosing a numeraire output, and this choice will generally influence estimates of the parameters of both the profit frontier and technical inefficiency. In the following, a measure of profit efficiency will be obtained, conditional on the choice of a numeraire output, and then decomposed into separate sources of profit inefficiencies.

3.2 HOMOGENEITY OF TECHNOLOGY AND SEPARABLE PROFIT FRONTIER

Let the technology of a firm be represented by a transformation function, \( T(y, x) \leq 0 \), where \( y \) and \( x \) are vectors of outputs and inputs, respectively. Assume that \( T(\cdot) \) is continuous and twice differentiable in the nonnegative orthant, and that \( T(\cdot) \) is strictly increasing in each element of \( y \) and strictly decreasing in each element of \( x \). Then, \( T(y, x) = 0 \) is referred to as a transformation frontier and \( T(y, x) < 0 \) indicates that the input-output combination of the firm is located inside the transformation frontier. The distance to the frontier is referred to as technical inefficiency, and can be measured in an output direction or an input direction.

Let \( e^{-u} \) be the distance to the transformation frontier measured in an output direction. Then \( e^{-u} \) measures output oriented technical inefficiency so that technical efficiency (inefficiency) is indicated by \( u = 0 \) (\( u > 0 \)). Given \( e^{-u} > 0 \), or the presence of technical inefficiency in production, \( T(y, x) < 0 \), can be represented equivalently as \( T(ye^{-u}, x) = 0 \). Given output prices \( p \) and input prices \( w \), the actual profit of the firm can be written as

\[
\pi^u(p, w, u) = \max_{y, x} \left\{ p \cdot y - w \cdot x; \ T(ye^{-u}, x) = 0 \right\}.
\]
This profit function is related to maximum profit \( \pi(p, w) \equiv \text{Max}_{y,x} \left\{ p \cdot y - w \cdot x; \ T(y, x) \leq 0 \right\} \) as

\[
\pi^u(p, w, u) \equiv \pi(pe^u, w). \tag{2}
\]

Clearly, \( \pi(pe^u, w) \) is linearly homogeneous in \((pe^u, w)\).

Whereas linear homogeneity of the cost frontier in \(we^u\) enables input oriented technical inefficiency \(e^{-u}\) to be separated from \(w\), since \(C(y, w_i e^*, w_i e^*) = w_i e^* C(y, w_i / w_i)\), linear homogeneity of the profit frontier in \((pe^u, w)\) is of no use in separating unobserved technical inefficiency from the vector of prices. An attempt to separate out \(e^u\) from \(p\) would lead to \(we^{-u}\) as an argument of the profit frontier. When linear homogeneity of profit frontier is useless, estimating a translog form of the profit frontier includes both the quadratic and linear forms of unobserved output oriented technical inefficiency, \(e^{-u}\). When panel data are available, the conventional treatment (Schmidt and Sickles, 1984), in which technical inefficiency is specified as an unobserved individual effect, is feasible only with a SUR model that drops the profit frontier. Kumbhakar et al. (2001), using data on Spanish savings banks, illustrated this approach, but the price of dropping the profit frontier is a loss of precision in the estimates of parameters.

However, dropping the profit frontier does not provide a complete solution for the estimation of the profit system if allocative inefficiencies are allowed. The empirical difficulty is compounded because technical inefficiency is attached to the composite form of shares, which is required to link the shadow profit to the actual profit.\(^{13}\) In addition, since unobserved technical inefficiency appears in the share equations, as Kumbhakar and Lovell (2000, p 252) point out, omission of any share equation will lead to biased and inconsistent estimates of all parameters.

\(^{13}\) See equation (5) below. Technical inefficiency appears in the bracket of (5) and in \(\pi^s(p^s e^*, w^s)\) as a quadratic form.
These difficulties, however, vanish if homogeneity of technology is imposed to separate unobserved technical inefficiency from the profit frontier. The following definition and propositions show that, assuming homogeneity in technology, the profit frontier is separable into output-oriented technical inefficiency and the normalized profit function.

**Definition** A firm’s normalized profit function is defined as

\[
\hat{\pi}(\hat{p}, w) \equiv \max_{y,x} \left\{ y_N + \sum_{m=1}^{M} \hat{p}_m y_m - \sum_{j=1}^{J} w_j x_j; \ T(y, x) = 0 \right\} \tag{3a}
\]

where \( y_N \) is numeraire output, \( y = [y_N, y_m] \), \( \hat{p} = p / p_N \), and \( \hat{p}_m = p_m / p_N \).

A firm’s normalized profit function has the following characteristics. First, \( \hat{\pi}(p, w) \) is convex in \( (p, w) \). Second, \( \hat{\pi}(p, w) \) is homogeneous of degree \( r/(r-1) \) in \( w \) if and only if \( T(y, x) = 0 \) is almost homogeneous of degree \( (r, 1) \) in \( (y, x) \)\(^{14}\).

When homogeneity of the technology is assumed, the normalized profit function is utilized, the profit frontier is separable into output-oriented technical inefficiency and the normalized profit function. Proposition 1 states the necessary and sufficient conditions for this separability. The proof is provided in the appendix D.

**Proposition 1** \( \pi(p e^w, w) = (p_N e^w)^{1-r} \hat{\pi}(\hat{p}, w) \) if and only if \( T(y, x) = 0 \) is almost homogeneous of degree \( (r, 1) \) in \( (y, x) \).

\(^{14}\) Following Lau (1973), almost homogeneous of degree \( (r, 1) \) in \( (y, x) \) is defined by \( T(k'y, kx) = kT(y, x) \), and a proof of the second characteristic is found in Lau (1973).
When the profit frontier is separable, a set of parametric restrictions on the normalized profit function needs to be imposed in the estimation. It turns out that homogeneity of degree \( r/(r-1) \) of the normalized profit function in \((\hat{p}, w)\) is equivalent to linear homogeneity of the profit frontier in \((pe^w, w)\). Proposition 2 states this equivalence. The proof is provided in the appendix D.

**Proposition 2** Assume \( T(y, x) = 0 \) is almost homogeneous of degree \((r, 1)\) in \((y, x)\). Linear homogeneity of \( \pi(pe^w, w) \) in \((pe^w, w)\) is equivalent to \( \hat{\pi}(\hat{p}, w) \) being homogenous of degree \( r/(r-1) \) in \( w \).

Applying Hotelling’s lemma, revenue and cost-share equations are easily derived as in the case of the profit function. But the revenue share of numeraire output can only be derived from the other shares:

\[
\hat{R}_m(\hat{p}, w) = \frac{\partial \ln \hat{\pi}}{\partial \ln \hat{P}_m} = \frac{y_i \hat{P}_m}{\hat{\pi}} \quad \text{for } m = 1, \ldots, M \tag{3b}
\]

\[
\hat{S}_j(\hat{p}, w) = \frac{\partial \ln \hat{\pi}}{\partial \ln w_j} = -\frac{x_j w_j}{\hat{\pi}} \quad \text{for } j = 1, \ldots, L \tag{3c}
\]

\[
\hat{R}_N(\hat{p}, w) = 1 - \sum_{m=1}^{M} \hat{R}_m(\hat{p}, w) - \sum_{j=1}^{L} \hat{S}_j(\hat{p}, w). \tag{3d}
\]
3.3 THE PROFIT SYSTEM WITH TECHNICAL AND ALLOCATIVE INEFFICIENCIES

Following the shadow-price approach, initiated by Lau and Yotopoulos (1971) and developed further by Atkinson and Halvorsen (1980), the profit frontier in equation (2) is parallel to an arbitrarily constructed shadow profit frontier

\[ \pi(p^S e^u, w^S) \equiv \text{Max}_{y,x} \{ p^S e^u \cdot y - w^S \cdot x; T(y, x) = 0 \}, \]  

where \( p^S \) and \( w^S \) are the shadow prices of outputs and inputs, respectively. Shadow prices are also constructed from market prices and their corresponding degree of allocative inefficiencies. That is, \( p_m^S = \theta_m p_m \) and \( w_i^S = \theta_i w_i \) are the shadow prices of output \( m \) and input \( l \), respectively.

Then, the shadow profit frontier is linked to actual profit \( \pi^a \) as

\[ \pi^a = \pi(p^S e^u, w^S) \left( \frac{R_N^S}{\theta_N} + \sum_{m=1}^{M} \frac{R_m^S}{\theta_m} + \sum_{l=1}^{L} \frac{S_l^S}{\theta_l} \right) \]  

where \( R_N^S, R_m^S \) and \( S_l^S \) are the revenue and cost shares of the shadow-profit frontier so that, for example, \( R_m^S = \frac{y_m e^u p_m}{\pi(p^S e^u, w^S)} \).

Under the homogeneity assumption in Proposition 1, the shadow profit frontier (4) is separable. That is,

\[ \pi(p^S e^u, w^S) \equiv \left( \frac{p_N^S e^u}{\hat{p}^S, w^S} \right)^{1-r} \hat{\pi}(p^S, w^S), \]  

where \( \hat{p}^S = p^S / p_N^S \).

Combining (5) and (6), the actual profit function can be written as

\[ \pi^a = \left( p_N^S e^u \right)^{1-r} \hat{\pi}(\hat{p}^S, w^S) \left( \frac{R_N^S}{\theta_N} + \sum_{m=1}^{M} \frac{R_m^S}{\theta_m} + \sum_{l=1}^{L} \frac{S_l^S}{\theta_l} \right), \]  

(7a)
The actual revenue and cost share equations are obtained using identity (2) and equation (5). That is,

\[ R_m^a = \frac{y_m e^u p_m}{\pi^a} = \left( \frac{R^S_m}{\theta_m} \right) \left( \frac{R^S_N}{\theta_N} + \sum_{m=1}^M \frac{R^S_m}{\theta_m} + \sum_{l=1}^L S^S_l \right)^{-1} \quad \text{for } m = 1, \cdots, M \]  

(7b)

\[ S_l^a = \frac{w_l x_l}{\pi^a} = \left( \frac{S^S_l}{\theta_l} \right) \left( \frac{R^S_N}{\theta_N} + \sum_{m=1}^M \frac{R^S_m}{\theta_m} + \sum_{l=1}^L S^S_l \right)^{-1} \quad \text{for } l = 1, \cdots, L \]  

(7c)

\[ R_N^S = \frac{y_N e^u p_N}{\pi^a} = \left( \frac{R^S_N}{\theta_N} \right) \left( \frac{R^S_N}{\theta_N} + \sum_{m=1}^M \frac{R^S_m}{\theta_m} + \sum_{l=1}^L S^S_l \right)^{-1}. \]  

(7d)

The above profit system can be implemented empirically when the normalized profit function is given the translog form

\[
\ln \hat{\pi}(\hat{p}^S, w^S) = \alpha_n + \sum_{m=1}^M \alpha_m \ln \hat{p}^S_m + \sum_{l=1}^L \beta_l \ln w^S_l \\
+ \sum_{m=1}^M \sum_{l=1}^L \eta_{ml} \ln \hat{p}^S_m \ln w^S_l \\
+ \frac{1}{2} \left[ \sum_{m=1}^M \sum_{k=1}^M \alpha_{mk} \ln \hat{p}^S_m \ln \hat{p}^S_k + \sum_{l=1}^L \sum_{j=1}^L \beta_{lj} \ln w^S_l \ln w^S_j \right],
\]

(8)

where the symmetry restrictions \( \alpha_{mk} = \alpha_{km}, \forall m, k \) and \( \beta_{lj} = \beta_{jl}, \forall j, l \) are imposed. The shadow profit system is then

\[
\ln \pi(p^S e^u, w^S) = \frac{1}{1-r} \left( \ln p^S_N + u \right) + \ln \hat{\pi}(\hat{p}^S, w^S)
\]

(9a)

with the shadow share equations for the normalized profit function given by

\[ \hat{R}_m^S = \alpha_m + \sum_{l=1}^L \eta_{ml} \ln w^S_l + \sum_{k=1}^M \alpha_{mk} \ln \hat{p}^S_k \quad \text{for } m = 1, \cdots, M \]  

(9b)

\[ \hat{S}_l^S = \beta_l + \sum_{m=1}^M \eta_{ml} \ln (\hat{p}^S_m) + \sum_{j=1}^L \beta_{lj} \ln w^S_j \quad \text{for } l = 1, \cdots, L. \]  

(9c)
\[ \hat{R}_N^S = 1 - \sum_{m=1}^{M} \hat{R}_m^S + \sum_{l=1}^{L} \hat{S}_l^S. \]  

(9d)

Once the shadow profit system is specified, it is trivial to specify the actual profit system (7a)-(7d).

Since parametric restrictions are imposed on the profit frontier \( \ln \pi(p^S e^u, w^S) \), because \( \pi(p^S e^u, w^S) \) is linearly homogenous in \( (p^S e^u, w^S) \), parametric restrictions are also imposed on the normalized profit function \( \ln \hat{\pi}(\hat{p}^S, w^S) \) because \( \hat{\pi}(\hat{p}^S, w^S) \) is homogenous of degree \( \frac{r}{r-1} \) in \( w \). The parametric restrictions on \( \ln \hat{\pi}(\hat{p}^S, w^S) \) turn out to be equivalent to the parametric restrictions on \( \ln \pi(p^S e^u, w^S) \), as Proposition 2 states. Thus, the parametric restrictions on \( \ln \hat{\pi}(\hat{p}^S, w^S) \) are the same as these on \( \ln \pi(p^S e^u, w^S) \) except that parameters on numeraire output \( p^S_p \) in \( \ln \pi(p^S e^u, w^S) \) to zero as if \( \ln \pi(p^S e^u, w^S) \) is specified as a translog form.

The parametric restrictions on \( \ln \hat{\pi}(\hat{p}^S, w^S) \) are

\[
\sum_{m=1}^{M} \alpha_m + \sum_{l=1}^{L} \beta_l = 1,
\]

\[
\sum_{k=1}^{M} \alpha_{mk} = \sum_{l=1}^{L} \eta_{ml} = 0 \ \forall \ m, \ \sum_{j=1}^{L} \beta_{lj} = \sum_{m=1}^{M} \eta_{ml} = 0 \ \forall \ l, \quad (9e)
\]

\[
\sum_{m=1}^{M} \sum_{k=1}^{M} \alpha_{mk} = \sum_{l=1}^{L} \sum_{j=1}^{L} \beta_{lj} = \sum_{m=1}^{M} \sum_{l=1}^{L} \eta_{ml} = 0 \ \forall \ m, k, l, j.
\]

Shadow price elasticities of outputs and inputs are calculated by using output supplies, \( y(p^S e^u, w^S) = R^S \pi / p^S \), and input demands, \( x(p^S e^u, w^S) = -S^S \pi / w^S \). Here, we present only the price elasticities of numeraire output because the others are trivially derived.\(^{15}\)

\(^{15}\) For details, see Kumbhakar (2001, p 6).
3.4 DECOMPOSITION OF PROFIT EFFICIENCY

Profit efficiency is defined as the ratio of actual profit to maximum profit associated with best practice, and is bounded above by unity. Since technical and allocative inefficiencies lower profit, profit efficiency can be decomposed to reflect the sources of these two inefficiencies.

Even after accounting for both technical and allocative inefficiencies, it is possible that a firm produces at the wrong scale. To capture scale inefficiency, profit efficiency ($\pi E$) is decomposed into four parts: profit technical inefficiency ($\pi TE$), profit allocative inefficiency in the input mix ($\pi AEI$), profit allocative inefficiency in the output mix ($\pi AEO$), and profit scale inefficiency ($\pi SE$).

$$
\pi E = \frac{\pi^a}{\pi^{\theta=1}} \leq 1
$$

$$
= \frac{\pi^a}{\pi^{\theta=1}} \frac{\pi^a}{\pi^{\theta=1}} \frac{\pi^a}{\pi^{\theta=1}} \frac{\pi^a}{\pi^{\theta=1}}
$$

$$
= \pi TE \times \pi AEI \times \pi AEO \times \pi SE
$$
where \( u = 0 \) indicates no technical inefficiency, and \( \theta_m = 1 \) and \( \theta_l = 1 \) indicate no allocative inefficiency in the input mix and in the output mix, respectively. Thus, \( \pi^a\big|_{u=0, \theta=1} \) is the maximum profit that the firm could earn.

Two points must be addressed in this decomposition of profit efficiency. First, profit technical inefficiency is a multiple of output technical inefficiency (\( u \)) since technology is homogeneous. That is, \( \ln \pi_{TE} = \left(1/(1-r)\right)u \). Second, profit scale efficiency as defined here can be greater than, equal to, or less than unity. This becomes obvious when the expression for profit scale efficiency is rearranged to be

\[
\pi_{SE} = \frac{\left( \left. \pi^a \right|_{u=0, \theta_x=1} \right)}{\left( \left. \pi^a \right|_{u=0, \theta_x=1} \right)} = \frac{\left. \pi_{AEO} \right|_{\theta_l=1}}{\left. \pi_{AEO} \right|_{\theta_l=1}}
\]

(11b)

\[
= \frac{\left( \left. \pi^a \right|_{u=0, \theta_x=1} \right)}{\left( \left. \pi^a \right|_{u=0, \theta_x=1} \right)} = \frac{\left. \pi_{AEI} \right|_{\theta_u=1}}{\left. \pi_{AEI} \right|_{\theta_u=1}}.
\]

(11c)

That is, profit scale efficiency is larger than 1 if there is not much room for improvement in profit allocative efficiency in the output mix as the incorrect input mix improves. Or equivalently, profit scale efficiency is larger than 1 if the improvement in profit allocative efficiency in the input mix is small as the incorrect output mix improves.\(^{16}\)

This decomposition of profit efficiency is built upon the benchmark technology in place at the current time. Hence, if productivity change is of importance, changes in the benchmark technology should be considered. Taking account of this, productivity at time \( t \) is defined as

---

\(^{16}\) Kumbhakar and Lovell (2000, pp 57-60) provide a complete theoretical explanation of the decomposition of profit efficiency.
productivity at 
\[ t = \frac{\pi E'(p^S_{t+1}e^u, w^S_{t+1})}{\pi E'(p^S_t e^u, w^S_t)} \]
\[ = \left( \frac{\pi E'^{t+1}(p^S_{t+1}e^u, w^S_{t+1})}{\pi E'(p^S_t e^u, w^S_t)} \right) \left( \frac{\pi E'(p^S_{t+1}e^u, w^S_{t+1})}{\pi E'(p^S_{t+1}e^u, w^S_{t+1})} \right), \]

where superscripts \( t \) and \( t+1 \) denote time \( t \) and \( t+1 \), respectively.

Given this definition of productivity, changes in productivity (\( \Delta \)productivity) are attributable to two sources: changes in profit efficiency (\( \Delta \pi E \)) and changes in technology (\( \Delta T \)).

\[ \Delta \text{productivity} = \ln \frac{\pi E'^{t+1}(p^S_{t+1}e^u, w^S_{t+1})}{\pi E'(p^S_t e^u, w^S_t)} - \ln \frac{\pi E'^{t+1}(p^S_{t+1}e^u, w^S_{t+1})}{\pi E'(p^S_{t+1}e^u, w^S_{t+1})} \]
\[ = \Delta \pi E + \Delta T. \]

3.5 EMPIRICAL APPLICATION

3.5.1 DATA AND ESTIMATION

A profit system with technical and allocative inefficiencies is estimated using data from Korean savings banks. Korean savings banks are similar to U.S. savings and loan associations (S&L). Thus, the sources and uses of funds are much simpler than in commercial banks. The main sources of funds are deposits, borrowed funds, and capital while the main use of funds are loans and other securities. But, unlike an S&L, mortgages are rarely held by Korean savings banks.

A bank is a financial intermediary which transforms various financial and physical resources into loans and investments. As Sealey and Lindley (1977) pointed out, the failure to consider the financial intermediation function of banks has led researchers to identify outputs and inputs incorrectly and to analyze the technical aspects of production and cost incorrectly. This point has
led to the development of the financial intermediation approach that is now most commonly followed in this literature.

Following the financial intermediation approach, the variables are defined in the following way. First, the input variables are labor \( x_1 \) and deposits \( x_2 \). Since borrowed funds are a relatively small fraction of liabilities, they are included in deposits. The corresponding prices of the inputs are constructed as follows: the price of labor \( w_1 \) is the sum of salaries and employment benefits divided by the number of employees; the price of deposits \( w_2 \) is the interest paid on deposits divided by the volume of deposits. Second, the output variable is defined to be loans \( y_1 \) and other non-financial assets \( y_2 \). Since cash and securities are typically a relatively small portion of a bank’s assets, they are combined with loans. Non-financial assets are classified as an output because some savings banks make profits from purchasing junk bonds. The price of loans \( p_1 \) is defined as interest earnings divided by loans, as is the price of non-financial assets \( p_2 \). Third, as Berger and Mester (2003) justified, various quasi-fixed inputs are introduced to control for the characteristics of the banking industry, such as: (i) off-balance sheet items \( z_1 \) that control for the quality of loans, on the assumption that credit risks increase as the size of the loan increases; (ii) financial capital \( z_2 \) to control for regulatory supervision, on the assumption that banks should meet regulatory capital requirements; and (iii) physical capital \( z_3 \) to control for the difficulty of measuring its price.

Finally, variable profit \( \pi \) is defined as revenue net of variable costs. This concept reflects

---

17 There is some controversy about the assumption.
18 To control for the quality of loans, other scholars use different variables; Hughes-Mester (1993) and Mester (1996) used non-performing loans while Berger et al. (1992) used loan-losses.
19 Hughes et al. (1996, 1997) and Hughes-Moon (1995) rejected the assumption of risk neutrality on which is based upon standard cost and profit efficiencies by arguing that some banks may hold a higher level of financial capital if they are more risk averse.
variable-profit maximization rather than long-run profit maximization because quasi-fixed inputs are included as controls, as explained above.

The profit maximization problem stated in (2) is changed to the short-run variable-profit maximization problem,

$$\pi(p^e, w, z) = \max_{y, x} \left( \pi(p^e) y - wx; \ T(y, x, z) = 0 \right). \quad (14)$$

The variable-profit maximization problem necessarily changes the functional form of (9a)–(9c), and thus the corresponding actual profit system, by including quasi-fixed input variables ($z$) as arguments of the normalized profit function, or $\hat{\pi}(p^s, w^s, z)$.

Data for this study are taken from information on Korean savings banks collected by the Korean Financial Supervisory Service (FSS). It is required that all savings banks submit annual reports to the FSS. We used data for the years 2002-2006, and selected a sample of 82 banks by deleting observations on banks with negative profits because the log of negative profits is not defined. All data are deflated by the GDP deflator, as is usually done in this literature, e.g., Berger and Mester (2003) and Wheelock and Wilson (1999). Table 3.1 provides the descriptive statistics on the variables describing the sample of Korean savings banks.

A prerequisite for estimating the profit system is empirical specification of technical and allocative inefficiencies, as well as changes in technology. Then, iterative feasible generalized least square (IFGLS) is used to estimate the profit system, comprised of the profit function, M revenue-share equations and L cost-share equations. In this case, 2 output 2 input equations, one revenue-share and two cost-share equations are estimated, along with the profit function.

First, the technical inefficiency of an individual firm is specified to be linear and time-varying. That is, the ith firm’s technical inefficiency at time $t$ is,

$$u_{it} = u_i (1 + \kappa_t) \quad \text{for } i = 1, \ldots, 82. \quad (15)$$
A more flexible form for technical inefficiency (i.e., $u_d = \kappa_{ai} + \kappa_{it}t + \kappa_{2t^2}$) may be employed, as introduced in Cornwell, Schmidt, and Sickles (1990). However, this form requires more parameters to be estimated. Thus, if $N$ is large and $T$ is small, the more flexible form is cumbersome. Of course, technical inefficiency may be assumed to be time-invariant (i.e., $u_t$). However, the longer the time span of the data, the less justified is the underlying assumption that technical inefficiency is constant over time.

Second, allocative inefficiency is specified to be linear and time-varying, as in Atkinson and Primont (2002). That is,

$$\theta_d = \theta_{d0} + \theta_{dt}t$$

for $d = N, Y, 1, 2$. (16)

It would be odd if allocative inefficiency is assumed to be time-invariant when technical inefficiency is assumed to be time-varying. It should be noted that allocative inefficiency is firm-average, not firm-specific, because the data in this study are in the form of a short panel (large $N$ and small $T$).

Third, following traditional econometric practice, changes in technology are proxied by a time trend. The shadow profit frontier is of the form

$$\ln \pi^s \left( p^s e^s, w^s, z, t \right) = \alpha_y \ln p^s_N + u + \ln \hat{\pi} \left( \hat{p}^s, w^s, z, t \right)$$

(17)

where

$$\ln \hat{\pi} \left( \hat{p}^s, w^s, z, t \right) = \alpha_y \ln \hat{p}^s + \sum_{i=1}^{2} \beta_i \ln w^s_i + \alpha_{yy} \ln \hat{p}^s + \sum_{i=1}^{2} \eta_i \ln \hat{p}^s \ln w^s_i$$

$$+ \frac{1}{2} \sum_{i=1}^{2} \sum_{j=1}^{2} \beta_{ij} \ln w^s_i \ln w^s_j + \sum_{h=1}^{3} \gamma_h \ln z_h + \frac{1}{2} \sum_{h=1}^{3} \sum_{f=1}^{3} \gamma_{hf} \ln z_h \ln z_f$$

$$+ \sum_{h=1}^{3} \rho_{yh} \ln \hat{p}^s \ln z_h + \sum_{i=1}^{3} \sum_{h=1}^{3} \tau^s_{ih} \ln w^s_i \ln z_h$$

---

20 Baltagi and Griffin (1979) provides a clear-cut explanation about the traditional specification of technical change in order to contrast it with their proposed specification of technical change.
\[ + \delta_1 t + \delta_2 t^2 + \delta_1 \ln \hat{p}_s + \sum_{i=1}^2 \delta_2 t \ln w_i + \sum_{h=1}^3 \delta_3 t \ln z_h. \]

with symmetry restrictions imposed on the \( \beta_{lj} \) and \( \gamma_{hf} \). Because \( z \) and \( t \) are included as arguments of the shadow profit frontier, additional parametric restrictions for the homogeneity of \(( \hat{p}_s, w_s )\),
\[ \sum_h (\rho_{h}\beta + \sum_i \tau_{ih}) = 0 \text{ and } \delta_i \gamma + \sum_i \delta_{2i} = 0, \]
need to be added to (9c). Then, changes in technology in (13) are represented as
\[ \Delta T = \delta_1 t + \delta_2 t + \delta_1 \ln \hat{p}_s + \sum_{i=1}^2 \delta_2 t \ln w_i + \sum_{h=1}^3 \delta_3 t \ln z_h. \]  

Once the shadow profit frontier is fully specified as in (17), then the shadow share equations are accordingly modified, as is the actual profit system using (7a)-(7d).

3.5.2 CHOICE OF NUMERAIRE AND EMPIRICAL RESULTS

As Lau (1978, p 170) clearly stated, there is no natural numeraire for a production function with multiple outputs. Therefore, a discussion on the choice of numeraire, which affects the empirical results, is warranted. Below, I will demonstrate the source of the discrepancy in the empirical results and discuss a rule of thumb for the choice of numeraire.

The discrepancy in the empirical results arising from the use of different numeraires is traced to two sources: an analytical source and an empirical source. The analytical source of the discrepancy comes from using a second-order approximation to the normalized profit function. That is, although Proposition 1 implies
\[ \pi(pe^\alpha, w) = \left( p_1 e^\alpha \right)^{\frac{1}{1-\tau}} \hat{\pi}(\hat{p}_2, w) = \left( p_2 e^\alpha \right)^{\frac{1}{1-\tau}} \hat{\pi}(\hat{p}_1, w), \]
the translog specification of \( \pi(pe^\alpha, w) = \left( p_1 e^\alpha \right)^{\frac{1}{1-\tau}} \hat{\pi}(\hat{p}_2, w) \) and \( \pi(pe^\alpha, w) = \left( p_2 e^\alpha \right)^{\frac{1}{1-\tau}} \hat{\pi}(\hat{p}_1, w) \)
yields \( (p_1 e^n)^{\frac{1}{1-\tau}} \hat{\pi}(\hat{p}_1, w) \neq (p_2 e^n)^{\frac{1}{1-\tau}} \hat{\pi}(\hat{p}_2, w) \). The empirical source of the discrepancy comes from the estimation technique, which is iterative feasible GLS. Since the profit function is analytically different in both normalizations, the disturbance covariance matrix \( \Omega \) of the multivariate normal distribution is different in the two models. It is sensible that the different \( \Omega \)'s lead to different parameter estimates. Note that, even in a seemingly unrelated regression (SUR) of the cost-share equations, parameter invariance does not hold for the one-step Zellner-efficient estimator if \( \Omega \) is changed by imposing symmetric restrictions on parameters (Berndt, 1990, p 474). Thus, the lack of equivalence in this study is different from that in Balk’s cost system. Lack of invariance in Balk’s cost system results from the fact that the inefficiency parameters are endogenous and, thus, iterative three stage least square is required (Kumbhakar and Karagiannis, 2004).

Lack of invariance with respect to choice of numeraire does not occur with production of a single output, and estimation employing homogeneity in technology can be fully utilized. However, the lack of invariance to choice of numeraire may be inevitable in models of production with multiple outputs. It is unfortunate to have to resort to a rule of thumb in that we most choose a numeraire which either fits previous studies or presumptive expectations from the data. In our application, for which no previous studies are available, careful analysis of the data leads us to choose \( y_2 \) as the numeraire.

Table 3.2 reports estimates of the coefficients from Model A and Model B in which the numeraires are \( y_1 \) and \( y_2 \), respectively. While estimates of the two models are obviously different, the differences are minimal. And, notably, the coefficients on numeraire output (\( \alpha_N \))

---

21 In order to make estimation easy, Balk (1997) proposed a cost system that sets the cost share weighted sum of the distortions equal to 1 on the assumption that shadow total cost equals actual total cost. Maietta (2002) reported that Balk’s cost system yields different estimates of parameters.
are 1.0371 for Model A and 1.0213 for Model B, which obey the theoretical restriction, \(1/(1 - r) > 1\). It is important that the empirical value of \(\alpha_n\) be larger than 1 because the assumption of homogeneity in technology is not testable. Since the profit system without assuming homogeneity in technology cannot be estimated, the parametric restriction of homogeneity in technology is not testable.

Table 3.3 shows time-varying technical and allocative inefficiencies estimated from Model A and Model B. Inefficiencies estimated for the two models are of different magnitude and trend alike. Firm-average technical inefficiency of Model A contrasts with that of Model B in its trend, as do allocative inefficiencies of outputs. Table 3.4 shows profit efficiency and its decomposition. Again, the two models provide contrasting results. The estimates for Model A indicate that the profit efficiencies of savings banks are very low and become worse over time, mainly due to deteriorating allocative efficiency of the output mix and scale efficiency. Plainly speaking, Model A indicates that profit efficiency became lower as banks depended relatively more on non-financial assets over loans in the course of moderately increasing interest rates. To the contrary, Model B shows that profit efficiencies improved due to soaring allocative efficiency of output mix and scale efficiency. Plainly speaking, Model B indicates that banks could make more profit over time because banks properly responded to increasing interest rates by increasing loans and that loans increased to the proper scale.

In both models, profit efficiency is lowered greatly due to technical inefficiency and a low level of allocative inefficiency in the input mix. That is, Korean savings banks are either costly in their financial intermediary functions or have low labor productivity, or both. In particular, it is implied that Korean savings banks are weakly responsive to an increase in wages.
Given the contrasting results from the two models, a close examination of the data is called for and is summarized below. First, as shown in the upper panel of Figure 3.1, the variable profits of savings banks doubled in 5 years. Profit efficiency estimated from Model B follows the trend in actual profit. Second, the price of loans increased after 2003 whereas the earnings from non-financial assets fell. As shown in the central panel of Figure 3.1, profit efficiency from Model B approximates the price of loans more closely than that from Model A. Third, the price of labor increased over time, whereas the price of deposits fell. The lower panel of Figure 3.1 shows that profit efficiency from Model B increases as the cost of borrowing money declines over time. Then, a rule of thumb is to resort to Model B rather than Model A.

The productivity change reported in Table 3.5 also shows the different results from the two models. Model B indicates that there has been a substantial gain in productivity during the full sample period, although it was diminishing over time, and which resulted from an increase in profit efficiency as well as positive technical change. The diminishing gain in productivity may imply that Korean savings banks responding to a few years of high profit either faced greater credit risk in loans or loosened tight management over employment and other expenses. To the contrary, Model A shows that both profit efficiency and technical change decreased such that productivity dropped over the period of 2003-2006.

Table 3.6 shows supply and demand elasticities. Own price elasticities in both models fall well into theoretical ranges; positive for supply and negative for demand. However, responsiveness to price change differs in the two models; savings banks respond relatively greater to $p_1$ and $\omega_1$ in Model A, while they respond relatively greater to $p_2$ and $\omega_2$ in Model B.
3.6 CONCLUSION

This paper proposes a way of estimating profit systems with technical and allocative inefficiencies in order to identify the sources of profit inefficiency. Prior attempts to estimate profit systems with technical and allocative inefficiencies have not been fully successful. The drawback to panel-data, fixed-effect estimation is that the profit function has a composite form of technical inefficiency in its nonlinear form. To overcome this drawback, an assumption of homogeneity in the technology is introduced to separate technical inefficiency from the profit frontier.

Assuming homogeneity of technology in a model with multiple outputs necessitates the choice of a numeraire, while the separability it introduces to the profit function makes estimation possible. The choice of numeraire poses a serious problem if the empirical results are highly sensitive to that choice.

In our application to Korean savings banks, two numeraires were employed and yield different profit efficiencies that were then decomposed into the sources of inefficiency (technical, allocative, and scale inefficiencies) and productivity change. Careful investigation of the data led us to conclude that estimates obtained using non-financial assets as the numeraire provides us with more reasonable measures of the sources of inefficiencies.
<table>
<thead>
<tr>
<th></th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>0.7844</td>
<td>0.3102</td>
<td>0.3133</td>
<td>0.4559</td>
<td>0.4467</td>
</tr>
<tr>
<td></td>
<td>(1.2382)</td>
<td>(0.4072)</td>
<td>(0.4958)</td>
<td>(1.2211)</td>
<td>(0.8286)</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.1069</td>
<td>0.1130</td>
<td>0.0953</td>
<td>0.0919</td>
<td>0.0932</td>
</tr>
<tr>
<td></td>
<td>(0.0304)</td>
<td>(0.0305)</td>
<td>(0.0210)</td>
<td>(0.0164)</td>
<td>(0.0194)</td>
</tr>
<tr>
<td>$w_i$</td>
<td>0.0284</td>
<td>0.0317</td>
<td>0.0339</td>
<td>0.0364</td>
<td>0.0418</td>
</tr>
<tr>
<td></td>
<td>(0.0099)</td>
<td>(0.0102)</td>
<td>(0.0151)</td>
<td>(0.0135)</td>
<td>(0.0219)</td>
</tr>
<tr>
<td>$w_2$</td>
<td>0.0648</td>
<td>0.0568</td>
<td>0.0509</td>
<td>0.0485</td>
<td>0.0443</td>
</tr>
<tr>
<td></td>
<td>(0.0079)</td>
<td>(0.0063)</td>
<td>(0.0064)</td>
<td>(0.0044)</td>
<td>(0.0044)</td>
</tr>
<tr>
<td>$z_1$</td>
<td>5,646</td>
<td>10,459</td>
<td>8,648</td>
<td>9,611</td>
<td>9,642</td>
</tr>
<tr>
<td></td>
<td>(8,103)</td>
<td>(11,091)</td>
<td>(12,783)</td>
<td>(16,643)</td>
<td>(14,782)</td>
</tr>
<tr>
<td>$z_2$</td>
<td>14,851</td>
<td>13,555</td>
<td>13,883</td>
<td>14,419</td>
<td>16,053</td>
</tr>
<tr>
<td></td>
<td>(14,767)</td>
<td>(14,926)</td>
<td>(16,819)</td>
<td>(16,579)</td>
<td>(18,079)</td>
</tr>
<tr>
<td>$z_3$</td>
<td>13,214</td>
<td>14,402</td>
<td>16,717</td>
<td>20,158</td>
<td>27,427</td>
</tr>
<tr>
<td></td>
<td>(11,709)</td>
<td>(12,162)</td>
<td>(15,320)</td>
<td>(20,805)</td>
<td>(31,282)</td>
</tr>
<tr>
<td>$\pi$</td>
<td>8,516</td>
<td>11,282</td>
<td>10,710</td>
<td>13,052</td>
<td>19,300</td>
</tr>
<tr>
<td></td>
<td>(13,313)</td>
<td>(14,591)</td>
<td>(14,472)</td>
<td>(17,377)</td>
<td>(26,946)</td>
</tr>
</tbody>
</table>

Note: Standard deviations are in the parentheses. $z$ and $\pi$ are expressed in million won.
Table 3.2  Estimated Coefficients

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model A (Numeraire: $y_1$)</th>
<th>Model B (Numeraire: $y_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_y$</td>
<td>$[\ln \hat{p}_y]$</td>
<td>0.0299 (0.5090)</td>
</tr>
<tr>
<td>$\alpha_N$</td>
<td>$[\ln p_{\text{Numeraire}}]$</td>
<td>1.0371 (0.0598) *</td>
</tr>
<tr>
<td>$\alpha_{Yy}$</td>
<td>$[\ln \hat{p}_y, \ln \hat{p}_y]$</td>
<td>0.0342 (0.0039) *</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$[\ln w_1]$</td>
<td>-0.0704 (0.1598)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$[\ln w_2]$</td>
<td>0.1116 (0.4624)</td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>$[\ln w_1 \ln w_1]$</td>
<td>-0.0165 (0.0063) *</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>$[\ln w_1 \ln w_2]$</td>
<td>0.0267 (0.0095) *</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>$[\ln w_2 \ln w_2]$</td>
<td>0.0047 (0.0085)</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>$[\ln z_1]$</td>
<td>-0.1100 (0.0282) *</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>$[\ln z_2]$</td>
<td>0.1449 (0.1542)</td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>$[\ln z_3]$</td>
<td>-1.0507 (0.3996) *</td>
</tr>
<tr>
<td>$\gamma_{11}$</td>
<td>$[\ln z_1 \ln z_1]$</td>
<td>0.0121 (0.0039) *</td>
</tr>
<tr>
<td>$\gamma_{12}$</td>
<td>$[\ln z_1 \ln z_2]$</td>
<td>-0.0046 (0.0026) **</td>
</tr>
<tr>
<td>$\gamma_{13}$</td>
<td>$[\ln z_1 \ln z_3]$</td>
<td>0.0099 (0.0037) *</td>
</tr>
<tr>
<td>$\gamma_{22}$</td>
<td>$[\ln z_2 \ln z_2]$</td>
<td>-0.0066 (0.0186)</td>
</tr>
<tr>
<td>$\gamma_{23}$</td>
<td>$[\ln z_2 \ln z_3]$</td>
<td>0.0032 (0.0260)</td>
</tr>
<tr>
<td>$\gamma_{33}$</td>
<td>$[\ln z_3 \ln z_3]$</td>
<td>0.1558 (0.0132) *</td>
</tr>
<tr>
<td>$\eta_{Y1}$</td>
<td>$[\ln \hat{p}_y \ln w_1]$</td>
<td>0.0003 (0.0006)</td>
</tr>
<tr>
<td>$\eta_{Y2}$</td>
<td>$[\ln \hat{p}_y \ln w_2]$</td>
<td>-0.0011 (0.0012)</td>
</tr>
<tr>
<td>$\rho_{Y1}$</td>
<td>$[\ln \hat{p}_y \ln z_1]$</td>
<td>0.0012 (0.0006) *</td>
</tr>
<tr>
<td>$\rho_{Y2}$</td>
<td>$[\ln \hat{p}_y \ln z_2]$</td>
<td>-0.0025 (0.0026)</td>
</tr>
<tr>
<td>$\rho_{Y3}$</td>
<td>$[\ln \hat{p}_y \ln z_3]$</td>
<td>0.0026 (0.0032)</td>
</tr>
<tr>
<td>$\tau_{11}$</td>
<td>$[\ln w_1 \ln z_1]$</td>
<td>-0.0011 (0.0004) *</td>
</tr>
<tr>
<td>$\tau_{12}$</td>
<td>$[\ln w_1 \ln z_2]$</td>
<td>0.0048 (0.0020) *</td>
</tr>
<tr>
<td>$\tau_{13}$</td>
<td>$[\ln w_1 \ln z_3]$</td>
<td>0.0055 (0.0024) *</td>
</tr>
<tr>
<td>$\tau_{21}$</td>
<td>$[\ln w_2 \ln z_1]$</td>
<td>0.0003 (0.0004)</td>
</tr>
<tr>
<td>$\tau_{22}$</td>
<td>$[\ln w_2 \ln z_2]$</td>
<td>-0.0047 (0.0021) *</td>
</tr>
<tr>
<td>$\tau_{23}$</td>
<td>$[\ln w_2 \ln z_3]$</td>
<td>0.0052 (0.0029) **</td>
</tr>
</tbody>
</table>

Note: Standard errors in the parentheses.
* denotes significance at the 0.05 level and ** denotes significance at the 0.1.
Table 3.2  Continued

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model A (Numeraire: $y_1$)</th>
<th>Model B (Numeraire: $y_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_1$</td>
<td>$[t]$</td>
<td>-0.4063 (1.6268)</td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>$[t^2]$</td>
<td>0.0312 (0.0127) *</td>
</tr>
<tr>
<td>$\delta_{1Y}$</td>
<td>$[t \ln \hat{p}_y]$</td>
<td>0.0000 (0.0022)</td>
</tr>
<tr>
<td>$\delta_{21}$</td>
<td>$[t \ln w_1]$</td>
<td>-0.0116 (0.0038) *</td>
</tr>
<tr>
<td>$\delta_{22}$</td>
<td>$[t \ln w_2]$</td>
<td>-0.0942 (0.0247) *</td>
</tr>
<tr>
<td>$\delta_{31}$</td>
<td>$[t \ln z_1]$</td>
<td>-0.0014 (0.0015)</td>
</tr>
<tr>
<td>$\delta_{32}$</td>
<td>$[t \ln z_2]$</td>
<td>-0.0060 (0.0066)</td>
</tr>
<tr>
<td>$\delta_{33}$</td>
<td>$[t \ln z_3]$</td>
<td>0.0107 (0.0087)</td>
</tr>
</tbody>
</table>

Note: Standard errors in the parentheses.
* denotes significance at the 0.05 level and ** denotes significance at the 0.1.
### Table 3.3  Time-varying Technical and Allocative Inefficiencies

<table>
<thead>
<tr>
<th>Year</th>
<th>$\theta_N$ (for $p_N$)</th>
<th>$\theta_r$ (for $\hat{p}_r$)</th>
<th>$\theta_1$ (for $w_1$)</th>
<th>$\theta_2$ (for $w_2$)</th>
<th>Technical Inefficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>0.7834</td>
<td>0.7862</td>
<td>0.2780</td>
<td>0.1610</td>
<td>0.3023</td>
</tr>
<tr>
<td>2003</td>
<td>1.0369</td>
<td>0.7771</td>
<td>0.3213</td>
<td>0.3043</td>
<td>0.3401</td>
</tr>
<tr>
<td>2004</td>
<td>1.2904</td>
<td>0.7680</td>
<td>0.3645</td>
<td>0.4475</td>
<td>0.3516</td>
</tr>
<tr>
<td>2005</td>
<td>1.5439</td>
<td>0.7589</td>
<td>0.4078</td>
<td>0.5907</td>
<td>0.3247</td>
</tr>
<tr>
<td>2006</td>
<td>1.7974</td>
<td>0.7498</td>
<td>0.4510</td>
<td>0.7339</td>
<td>0.2995</td>
</tr>
</tbody>
</table>

#### Model A (Numeraire: $y_1$)

<table>
<thead>
<tr>
<th>Year</th>
<th>$\theta_N$ (for $p_N$)</th>
<th>$\theta_r$ (for $\hat{p}_r$)</th>
<th>$\theta_1$ (for $w_1$)</th>
<th>$\theta_2$ (for $w_2$)</th>
<th>Technical Inefficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>0.4727</td>
<td>0.4909</td>
<td>0.1630</td>
<td>0.0945</td>
<td>0.2992</td>
</tr>
<tr>
<td>2003</td>
<td>0.4658</td>
<td>0.6403</td>
<td>0.1873</td>
<td>0.1767</td>
<td>0.3375</td>
</tr>
<tr>
<td>2004</td>
<td>0.4589</td>
<td>0.7897</td>
<td>0.2117</td>
<td>0.2589</td>
<td>0.3512</td>
</tr>
<tr>
<td>2005</td>
<td>0.4519</td>
<td>0.9392</td>
<td>0.2361</td>
<td>0.3411</td>
<td>0.3266</td>
</tr>
<tr>
<td>2006</td>
<td>0.4450</td>
<td>1.0886</td>
<td>0.2605</td>
<td>0.4233</td>
<td>0.3011</td>
</tr>
</tbody>
</table>

#### Model B (Numeraire: $y_2$)
<table>
<thead>
<tr>
<th>Year</th>
<th>Technical efficiency</th>
<th>Allocative efficiency in Input mix</th>
<th>Allocative efficiency in Output mix</th>
<th>Scale Efficiency</th>
<th>Profit Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>0.2909</td>
<td>0.1372</td>
<td>0.9371</td>
<td>8.7665</td>
<td>0.1421</td>
</tr>
<tr>
<td>2003</td>
<td>0.3286</td>
<td>0.2808</td>
<td>0.8591</td>
<td>2.0199</td>
<td>0.1604</td>
</tr>
<tr>
<td>2004</td>
<td>0.3400</td>
<td>0.3205</td>
<td>0.7797</td>
<td>1.6647</td>
<td>0.1414</td>
</tr>
<tr>
<td>2005</td>
<td>0.3131</td>
<td>0.3245</td>
<td>0.6999</td>
<td>1.4951</td>
<td>0.1072</td>
</tr>
<tr>
<td>2006</td>
<td>0.2880</td>
<td>0.3082</td>
<td>0.6212</td>
<td>1.4564</td>
<td>0.0805</td>
</tr>
</tbody>
</table>

Model B (Numeraire: y₂)

<table>
<thead>
<tr>
<th>Year</th>
<th>Technical efficiency</th>
<th>Allocative efficiency in Input mix</th>
<th>Allocative efficiency in Output mix</th>
<th>Scale Efficiency</th>
<th>Profit Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>0.2927</td>
<td>0.6403</td>
<td>0.4917</td>
<td>1.4614</td>
<td>0.1236</td>
</tr>
<tr>
<td>2003</td>
<td>0.3308</td>
<td>0.4720</td>
<td>0.5963</td>
<td>2.8043</td>
<td>0.2529</td>
</tr>
<tr>
<td>2004</td>
<td>0.3445</td>
<td>0.4325</td>
<td>0.6965</td>
<td>3.7542</td>
<td>0.3708</td>
</tr>
<tr>
<td>2005</td>
<td>0.3199</td>
<td>0.4220</td>
<td>0.7862</td>
<td>4.2648</td>
<td>0.4381</td>
</tr>
<tr>
<td>2006</td>
<td>0.2944</td>
<td>0.4284</td>
<td>0.8625</td>
<td>4.5724</td>
<td>0.4811</td>
</tr>
<tr>
<td>Year</td>
<td>EC</td>
<td>TC</td>
<td>PC</td>
<td>EC</td>
<td>TC</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>2002-2003</td>
<td>0.0525</td>
<td>0.0242</td>
<td>0.0767</td>
<td>0.3109</td>
<td>0.3245</td>
</tr>
<tr>
<td>2003-2004</td>
<td>-0.0548</td>
<td>-0.0156</td>
<td>-0.0703</td>
<td>0.1663</td>
<td>0.2537</td>
</tr>
<tr>
<td>2004-2005</td>
<td>-0.1200</td>
<td>-0.0624</td>
<td>-0.1824</td>
<td>0.0724</td>
<td>0.1726</td>
</tr>
<tr>
<td>2005-2006</td>
<td>-0.1246</td>
<td>-0.0977</td>
<td>-0.2223</td>
<td>0.0407</td>
<td>0.1004</td>
</tr>
</tbody>
</table>
Table 3.6 Supply and Demand Elasticities

<table>
<thead>
<tr>
<th></th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$x_1$</th>
<th>$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(Model A)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with respect to $p_1$</td>
<td>0.567</td>
<td>0.028</td>
<td>0.106</td>
<td>0.062</td>
</tr>
<tr>
<td>$p_2$</td>
<td>-0.259</td>
<td>1.068</td>
<td>1.202</td>
<td>1.246</td>
</tr>
<tr>
<td>$w_1$</td>
<td>-0.023</td>
<td>-0.039</td>
<td>-3.389</td>
<td>-0.178</td>
</tr>
<tr>
<td>$w_2$</td>
<td>-0.316</td>
<td>-0.268</td>
<td>3.540</td>
<td>-1.293</td>
</tr>
<tr>
<td><strong>(Model B)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with respect to $p_1$</td>
<td>0.019</td>
<td>0.752</td>
<td>0.991</td>
<td>0.979</td>
</tr>
<tr>
<td>$p_2$</td>
<td>0.252</td>
<td>5.920</td>
<td>0.280</td>
<td>0.293</td>
</tr>
<tr>
<td>$w_1$</td>
<td>-0.032</td>
<td>-0.029</td>
<td>-0.754</td>
<td>-0.195</td>
</tr>
<tr>
<td>$w_2$</td>
<td>-0.239</td>
<td>-0.245</td>
<td>-0.692</td>
<td>-1.268</td>
</tr>
</tbody>
</table>
Figure 3.1 Profit efficiencies and actual data

- **Actual profit**: The graph shows the actual profit over the years 2002 to 2006. The profit efficiency for bank A (pe_A) and bank B (pe_B) is plotted against the actual profit.

- **Price of loans**: The graph illustrates the price of loans for the same years. The efficiency of bank A (pe_A) and bank B (pe_B) is compared with the price of loans.

- **Price of deposits**: The graph depicts the price of deposits from 2002 to 2006. The efficiency of bank A (pe_A) and bank B (pe_B) is contrasted with the price of deposits.
CHAPTER 4

ALLOCATIVE INEFFICIENCY, SUBSTITUTABILITY, AND THE CHOICE OF A
QUANTITY INDEX

4.1 INTRODUCTION

This chapter addresses two questions. First, how does the degree of substitutability between
inputs affect the extent of allocative inefficiency? Second, how well is the degree of allocative
inefficiency captured by the Fisher and Tornqvist indices? Regarding the first question, we will
show that allocative inefficiency increases as the elasticity of substitution increases when a
technically efficient firm is operating with a CES production function with factor shadow prices
that differ from actual factor prices. Regarding the second question, we will show that an
increase in allocative inefficiency is accurately reflected in the Fisher quantity index but not in
the Tornqvist index, which may overstate or understate the deterioration in allocative efficiency.

Index numbers are important in many economic contexts. When we negotiate wage rates, we
evaluate changes in a consumer price index to measure how much the overall price level has
changed. When we evaluate the performance of an industry or a firm, we may wish to measure
total factor productivity (TFP). An index number is constructed whenever we aggregate or
summarize many different prices and/or quantities.

An important issue has been how an index can be constructed to reflect accurately the
economic phenomena of interest. To this end, various index formulae have been proposed and
widely used. In order to find a suitable index number, a number of desirable properties have been suggested and discussed. Fisher (1922), Eichhorn (1978), and Diewert (1992) proposed a number of intuitively appealing criteria that an index should satisfy. However, all of these studies recognized that no perfect index number is attainable22.

While the axiomatic approach does not resolve the index number problem, economic theory has been brought to bear in the hope of improving the choice among alternative indexes. Applying economic theory, the data observed for an index formula are assumed to be the result of optimizing behavior such as cost minimization or utility maximization. The Malmquist quantity index is a theoretic index that is the ratio of distance functions for two time periods when technology is represented by an unknown production function. If an underlying technology is specified and an index coincides with the theoretic index, the index is called exact. If the underlying technology is specified as a flexible functional form, e.g., translog, and an index approximates a theoretic index, the index is superlative. The importance of being superlative is that such an index is independent of the specific form of the underlying technology.

Diewert (1992) showed that the Fisher index is superlative under certain conditions, and Caves et al. (1982) showed that the Tornqvist index is superlative under similar conditions. These conditions all concern the specification of the flexible functional form of the unknown technology, and require that 1) the benchmark technology must exhibit constant returns to scale, 2) distance function must have a flexible functional form, 3) the period t and period t+1 distance functions must have identical second order coefficients, and 4) production in both periods must be allocatively efficient. The first three conditions are not terribly restrictive, but allocative

22 See the comment by Snyder (1923, p 418) on Fisher’s work, and see Proposition 1 in Diewert (1993, p 9) stating a result that is comparable to the Impossibility Theorem of Eichhorn (1978).
inefficiency may arise for many reasons. A failure to fulfill the final requirement is a serious shortcoming because it precludes the assumption of economic optimization.

Unfortunately, no study has investigated the degree to which a superlative index is compromised when allocative inefficiency exists. Therefore, this chapter will investigate this issue and derive its implications for the choice of an index. Diewert (1978, Appendix 2, Table 1) shows, using Canadian consumer data, that the Fisher and Tornqvist indices numerically approximate each other. Dumagan (2002) advocates using the Tornqvist index because it requires less data. The simulation results reported here indicate that the Fisher index provides a more accurate portrayal of deteriorating allocative inefficiency. The implication is that, when allocative inefficiency is detected, the Fisher index provides the more reliable indicator of economic growth.

In addition, we address the question of how allocative inefficiency varies when substitutability between factors changes because allocative inefficiency is closely related to the choice of technology. Section 2 examines the relation between allocative inefficiency and the substitutability between input factors, and shows that allocative inefficiency increases as the elasticity of substitution increases. Section 3 provides simulation results that indicate the impact of allocative inefficiency on the Fisher and Tornqvist indices, and draws implications from these results for the choice of an index. The final section offers concluding remarks.

4.2 ALLOCATIVE INEFFICIENCY AND FACTOR SUBSTITUTABILITY

Allocative inefficiency occurs when a firm fails to satisfy the first-order conditions for cost minimization, which requires that the marginal rates of technical substitution (MRTS) for each
pair of inputs equal the corresponding input price ratios. A firm may not satisfy the first-order
conditions for a variety of reasons, including government intervention as discussed by Averch
and Johnson (1962) in their analysis of rate-of-return regulation, and market power, which
distorts the terms of trade, as when employers must bargain with workers, either collectively in
unions, as analyzed by Rees (1963) and DeFina (1983), or through individual negotiations, as in
the model of intrafirm bargaining developed by Stole and Zwiebel (1996). In addition, the
presence of moral hazard may cause allocative inefficiency through shirking, as investigated by
Ross (1973) and Holmstrom (1979, 1982), or through opportunistic behavior, in which resources
are diverted to private consumption, as explored by Alchian and Demsetz (1972) and Williamson
(1979)\textsuperscript{23}.

The failure of a firm to satisfy the first-order conditions imposes a cost. This cost does not
disappear even if the firm is technically efficient, that is, if it operates on its production frontier.
As a reflection of this cost, allocative inefficiency is defined as the ratio of actual cost to
minimum cost under the assumption that technical efficiency is achieved, that is,

\[
\text{Allocative Inefficiency} (AI) \equiv \frac{\text{Actual Cost} \ (C^A)}{\text{Minimum Cost} \ (C^M)}.
\]

where minimum cost is the cost the firm would incur if its output were produced with an
allocatively efficient input mix\textsuperscript{24}.

The concept of allocative inefficiency stated above was originated by Farrel (1957). An
alternative way of capturing the degree of allocative inefficiency, referred to as the shadow price
approach, was originated by Lau-Yotopoulos (1971). The basic idea of this shadow price
approach is that allocative inefficiency arises when a firm behaves as if it is minimizing

\textsuperscript{23} Depken II, Redmount and Snow (2001) focused exclusively on shirking as the potential cause of allocative
inefficiency.

\textsuperscript{24} Allocative inefficiency is the reciprocal of allocative efficiency, the ratio of minimum cost to actual cost. For
details, see Kumbhakar and Lovell (2001, p 54).
production cost relative to perceived shadow factor prices, \( w^s \), that differ from actual factor prices, \( w \). The degree of distortion in the \( i \)th factor price is represented by \( k_i \) where \( w'_i = w_i k_i \).

Although the concept, measurement, and sources of allocative inefficiency have been widely and extensively studied, the relationship between allocative inefficiency and input substitutability has not been studied. It is surprising that we are not aware of how the magnitude of allocative inefficiency changes as underlying technology dictates substitutability between input factors. This relationship is important to any superlative index because a superlative index is valid only under certain conditions on the underlying technology. Without knowing this relation, we do not know whether allocative inefficiency increases or decreases as substitutability between input factors changes. Accordingly, we would not know how much a specific index is distorted by allocative inefficiency.

Here, we address the following question. For given degrees of distortion in the perceived or shadow factor prices represented by a vector \( k \), how does allocative inefficiency vary with the degree of substitutability between two inputs? In order to answer this question, we exploit the following two lemmas. Proofs are provided in the appendix.

**Lemma 1**  
For a linear homogeneous, two-factor production function, actual cost increases with increased distortion in an input price, ceteris paribus, at a rate that is directly proportional to the elasticity of substitution.

**Lemma 2**  
For a CES production function, the rate at which actual cost increases with increased distortion increases as the substitutability between inputs increases.
That is, increased distortion in an input price necessarily increases actual cost under the ceteris paribus assumption.

Lemma 1 states that, when the production function is linearly homogeneous, i.e., exhibits constant return to scale, the rate of increase in cost is proportional to the elasticity of substitution, \( \sigma \). Since the rate of cost increase also depends on the actual factor choices, Lemma 2 imposes the further restriction of CES technology to show that the rate of increase in actual cost increases with \( \sigma \). The intuition is illustrated in Figure 4.1. The dotted line is associated with the actual price ratio. Given the degree of distortion, \( k \), the MRTS associated with the shadow-price ratio is shown by the solid line at point B resulting from limited substitution between inputs. The MRTS at point B is larger than the MRTS at point A as a result of the stronger substitution between inputs. Hence, the same degree of distortion must lead to higher actual cost in the case of stronger substitution.

With the aid of Lemma 1 and Lemma 2, we establish the following proposition, which states that allocative inefficiency increases with the degree of substitutability between factors, ceteris paribus. Proof is provided in the appendix.

**Proposition 1** For a CES production function, allocative inefficiency increases as the substitutability between factors increases.
4.3 SIMULATIONS

In this section, we address the impact of allocative inefficiency and the substitutability between input factors on Tornqvist and Fisher quantity indices. For this purpose, simulations are conducted in two different rounds:

(1) Simulations under CES technology;

(2) Simulations under an unknown technology.

In the first round, we conduct simulations on the assumption that we know that the firm’s technology is CES in order to investigate how $\sigma$ affects allocative inefficiency as measured by the Tornqvist and Fisher indices. In the second round, we drop that assumption and generate data in order to investigate how much the two indices deteriorate as allocative inefficiency increases. With these simulations, we draw implications for the choice of an index.

4.3.1 DESIGN OF THE EXPERIMENT

4.3.1.1 SIMULATIONS UNDER CES TECHNOLOGY

In this first-round experiment, we assume that we know a firm’s technology. The advantages of this assumption are enormous since, with it, we can easily construct an allocatively inefficient input mix as well as a cost-minimizing input mix and calculate minimum cost. The greater advantage is, however, that we can control the degree of substitutability of inputs to see how well the indices perform as allocative inefficiency changes.
Consider the CES technology \( y = \left( \alpha_1 x_1^\rho + \alpha_2 x_2^\rho \right)^{1/\rho} \), where \( \alpha_1 + \alpha_2 = 1 \) and \( \rho = 1 - \frac{1}{\sigma} \). For this technology, the cost-minimizing input mix \((x)\) and minimum cost \((C^M)\) are respectively,

\[
x_j = y \left( \frac{\alpha_1}{\alpha_j} w_1^{-\rho} + \frac{\alpha_2}{\alpha_j} w_2^{-\rho} \right)^{1/\rho} \text{ for } j=1, 2, \quad (2a)
\]

\[
C^M = y \left( \frac{\alpha_1}{\alpha_1} w_1^{-\rho} + \frac{\alpha_2}{\alpha_2} w_2^{-\rho} \right)^{1-\rho/\rho}. \quad (2b)
\]

Then, an allocatively inefficient input \((\hat{x}_j)\) is easily constructed as we control \(\sigma\). The corresponding input \((\hat{x}_j)\) and actual cost \((C^A)\) are respectively

\[
\hat{x}_j = \left( y^\rho - \alpha_j \hat{x}_j \right)^{1/\rho} / \alpha_j \text{ for } j \neq i = 1, 2, \quad (3a)
\]

\[
C^A = w_1 \hat{x}_1 + w_2 \hat{x}_2. \quad (3b)
\]

One beauty of knowing technology is that shadow input prices \((\hat{w})\) are easily obtained using a distance function, \(D(x, y) = \left[ \alpha_1 x_1^\rho + \alpha_2 x_2^\rho \right]^{1/\rho} / y\), at \(\hat{x}\);

\[
\hat{w}_j = \frac{\partial D}{\partial x_j} = \frac{\alpha_j \hat{x}_j \left[ \alpha_1 (\hat{x}_1)^\rho + \alpha_2 (\hat{x}_2)^\rho \right]^{1/\rho}}{y} \text{ for } j = 1, 2. \quad (3c)
\]

With this information, we are able to calculate AI, given degree of distortion in input prices as we control \(\sigma\).

We are now ready to calculate two popular quantity indices to see how they perform when allocative inefficiency in the input mix exists. For a bilateral index which compares only two time periods, we assume that allocative inefficiency occurs at time \(t+1\). That is, the input mix at time \(t\) is cost-minimizing and allocatively efficient, whereas the input mix at time \(t+1\) is
allocatively inefficient. Note that observed inputs $\hat{x}$ are allocatively inefficient, whereas the cost-minimizing inputs $x$ are not observed if allocative inefficiency exists. Therefore, the Tornqvist quantity index (TQI) is calculated as

$$TQI = \prod_{j=1}^{K} \left( \frac{x_{j}^{t+1} - x_{j}^{t}}{x_{j}^{t}} \right)^{\frac{S_{j}^{t+1} - S_{j}^{t}}{2}} = \prod_{j=1}^{K} \left( \frac{\hat{x}_{j}^{t+1} - \hat{x}_{j}^{t}}{x_{j}^{t}} \right)^{\frac{\hat{S}_{j}^{t+1} - \hat{S}_{j}^{t}}{2}}$$

(4a)

where $S_{j}^{t}$ and $S_{j}^{t+1}$ are respectively cost shares of input $j$ at $t$ and $t+1$. $\hat{S}_{j}$ denotes the actual cost share of input $j$ when allocative inefficiency exists. The Fisher quantity index (FQI) is calculated as the geometric mean of a Laspeyres quantity index ($Q_{L}$) and a Paasche index ($Q_{P}$);

$$FQI = \left( Q_{L} \cdot Q_{P} \right)^{\frac{1}{2}}$$

(4b)

where $$Q_{L} = \frac{\sum_{j=1}^{K} w_{j} x_{j}^{t+1}}{\sum_{j=1}^{K} w_{j} x_{j}^{t}}$$ and $$Q_{P} = \frac{\sum_{j=1}^{K} w_{j}^{s+1} x_{j}^{t+1}}{\sum_{j=1}^{K} w_{j}^{s+1} x_{j}^{t}}.$$ From (4a) and (4b), it is inferred that the Fisher quantity index is identical to allocative inefficiency (AI) while the Tornqvist quantity index is not. When we take the log of (4a), we have

$$\log(TQI) = \sum_{j=1}^{K} \left( \frac{S_{j} + \hat{S}_{j}}{2} \right) \log \left( \frac{\hat{x}_{j}}{x_{j}} \right).$$

(4c)

That is, the log of the Tornqvist index is simply the weighted average of the growth rates of each input factor. This implies that if a weighted cost share has large portion of AI, this large portion of AI will be reflected in the Tornqvist index. Thus, the Tornqvist index suffers from Simpson’s paradox\(^{25}\), in which the weighted average of two events may be amalgamated.

\(^{25}\) For example, suppose that player A has higher batting averages than player B for both two different years. It is possible for player A’s weighted two-year batting average to be lower than a player B’s.
4.3.1.2 SIMULATIONS UNDER UNKNOWN TECHNOLOGY

In this second-round experiment, we drop the assumption that we know the technology. Dropping this assumption requires us to estimate the technology as well as recover the unobserved input quantities corresponding to the observed input quantities.

First, we estimate a flexible form for the input distance function given data on observed inputs \( \hat{x} \) and observed output \( y \), given by the translog function

\[
\ln D = \alpha_0 + \alpha_y \ln y + \sum_{i=1}^{K} \beta_i \ln \hat{x}_i + \frac{1}{2} \alpha_{yy} \ln y \ln y
\]
\[
+ \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} \beta_{ij} \ln \hat{x}_i \ln \hat{x}_j + \sum_{i=1}^{K} \delta_{iy} \ln \hat{x}_i \ln y + v,
\]

where \( v \) is statistical noise that is assumed to be distributed normally. The estimation of input distance function requires that we impose linear homogeneity in the inputs

\[
\left( \sum_{i=1}^{2} \beta_i = 1, \sum_{j=1}^{2} \beta_j = 2 \beta_i = \sum_{i=1}^{2} \delta_i = 0 \right)
\]
as well as the symmetry restrictions \( \beta_{ij} = \beta_{ji} \forall i,j \).

However, this distance function can not be estimated because we do not observe the dependent variable, \( \ln D \). Hence, we use linear homogeneity in the inputs to obtain

\[
\ln D \left( \frac{\hat{x}}{\hat{x}_j}, y \right) = - \ln \hat{x}_j - \ln D(\hat{x}, y).
\]

Setting \( u = \ln D(\hat{x}, y) \) and adding statistical noise \( (v) \) yields the empirically tractable distance function,

\[
- \ln \hat{x}_j = \ln D \left( \frac{\hat{x}}{\hat{x}_j}, y \right) - u + v
\]

26 Atkinson-Primont (2002) presented one way of estimating the distance function directly. That is, they specify the model \( 0 = \ln D(x, y) + \exp(v-u) \).
where \( \ln D(\bullet) \) is of the form given in (6a). Then, estimation of the distance function can be carried out either with the corrected ordinary least squares (COLS) method or any other frontier approach. Here, we use the COLS method for simplicity since we have cross-sectional rather than panel data. The COLS method adjusts the residuals obtained from OLS estimation as follows:

\[
\hat{u}_k = u_k - u_{\text{Max}},
\]

(7)

and defines \( TE = \exp(-\hat{u}_k) \). The COLS estimator is consistent \(^{27}\) (Greene, 1980b).

After estimating the technology, the unobserved cost-minimizing input quantities are recovered using the approach of Karagiannis, Kumbhakar, and Tsionas (2004). Using the estimated distance function, we obtain the system of equations;

\[
\begin{align*}
\frac{w_j x_j}{w_i x_i} &= \frac{\beta_j + \sum_{j=1}^{K-1} \beta_j \ln \left( \frac{x_j}{x_1} \right) + \delta_{jy} \ln y}{\beta_i + \sum_{j=1}^{K-1} \beta_j \ln \left( \frac{x_j}{x_1} \right) + \delta_{iy} \ln y} \\
\text{for } j = 2, \cdots, K.
\end{align*}
\]

(8)

Because (8) is non-linear, Newton’s method is used to obtain the \((K-1)\) factor ratios \( \left( x_j / x_1 \right) \).

Once the \((K-1)\) factor ratios \( \left( x_j / x_1 \right) \) are obtained, the numeraire input \( ( x_1 ) \) is obtained from (6b).

Finally, using the unobserved input quantities, we calculate allocative inefficiency as well as the Tornqvist and Fisher quantity indices shown in (4a) and (4b).

\(^{27}\) One possible drawback is that the COLS estimator is vulnerable to the presence of outliers, as the OLS estimation is. However, this possibility does not appear in our simulation because data generation follows the normal distribution.
In the first-round experiment, in which the technology is assumed to be known, we consider the simplest case:

\[ w_1 = w_2 = 1, \quad \alpha_1 = \alpha_2 = 0.5. \]

Given the input prices and technology parameters, the cost-minimizing input mix is \( x_1 = x_2 = 1 \) and minimum cost \( C^M \) is 2. For substitutability, we consider three alternative technologies in which the elasticities of substitution \( \sigma \) are, respectively,

\[ \sigma = 0.67, \quad \sigma = 1, \quad \sigma = 1.33. \]

For each \( \sigma \), an allocatively inefficient input mix \( \hat{x}_i \) is randomly generated to range from 0.5 to 1.5 along a uniform distribution. Then, following (3a)-(3c), the allocatively inefficient input mix \( \hat{x}_1, \hat{x}_2 \), actual cost \( C^A \), and the shadow input prices \( \hat{w}_1, \hat{w}_2 \) are constructed.

Figure 4.2 illustrates how allocative inefficiency \( AI \) changes with the shadow price ratio \( \hat{w}_2 / \hat{w}_1 \). Following (1) \( AI \), measured on the y-axis, shows the degree of allocative inefficiency. That is, 1 indicates allocative efficiency in which \( C^A = C^M = 2 \). 1.5 indicates allocative inefficiency in which \( C^A = 3 \) because \( C^M \) is fixed at 2. The shadow price ratio on the x-axis indicates the degree to which the input prices a firm perceives are distorted. That is, 1 indicates no distortion in a firm’s perceived input prices so that shadow prices are same as the actual prices. Any number other than 1 indicates distortion in a firm’s perceived input prices.

From Figure 4.2, we verify Proposition 1. That is, for each value of \( \sigma \), allocative inefficiency increases as the shadow price ratio deviates from 1, as actual cost increases, and as
prices are distorted by a constant ratio. Given the level of price distortion, allocative inefficiency increases as the elasticity of substitution between inputs increases.

Figure 4.3 depicts the Tornqvist and Fisher quantity indices for different increases in allocative inefficiency. Figure 3 verifies that the Fisher quantity index exactly approximates allocative inefficiency, and also shows that the Tornqvist quantity index understates the degree of allocative inefficiency. Although both indices increase with increased allocative inefficiency, the extent to which the Tornqvist index understates allocative inefficiency becomes larger as allocative inefficiency becomes larger. Therefore, the gap between the Fisher and Tornqvist quantity indices becomes larger as allocative inefficiency increases. This result accords with Dumacan (2002), who shows that the Tornqvist and Fisher quantity indices numerically approximate each other if the values of \( x_i \) and \( x_i^{t+1} \) are close for \( i = 1, \ldots, N \), (that is, a cost-minimizing input vector \( x_i \) is close to an allocatively inefficient vector \( x_i^{t+1} \) in our simulation.)

In the second-round experiment, where we drop the assumption of a known technology, we generated data randomly, recovered the unobserved cost-minimizing input vectors, and calculated both the Tornqvist and Fisher quantity indices. Here, we assume a one-output, three-input production technology because the method of Karagiannis, Kumbhakar, and Tsionas (2004) cannot be used for a two-input production technology. Under these assumptions, we generate data on three inputs \((\hat{x}_1, \hat{x}_2, \hat{x}_3)\), one output \((\hat{y})\), and three input prices \((w_1, w_2, w_3)\) that are i.i.d. \(N(\mu, 1)\). By the method explained in (4) to (7), we determine the cost-minimizing input quantities \((x_1, x_2, x_3)\).

Figure 4.4 shows that both indexes increase as allocative inefficiency increases, but at different rates. The x-axis shows average values of allocative inefficiency in the experiment, whereas the y-axis shows average values of the indices in the experiment. The Fisher quantity
index approximates allocative inefficiency more closely than the Tornqvist index. That is, the Tornqvist quantity index understates the degree of allocative inefficiency. Moreover, the extent to which the Tornqvist index understates aggravates as allocative inefficiency becomes larger.

4.4 CONCLUSION

We have shown that allocative inefficiency increases with input substitutability, that both the Tornqvist and Fisher indices increase as allocative inefficiency increases, and that the Fisher index captures increased allocative inefficiency more accurately than the Tornqvist index.

Diewert (1978) showed by numerical example that the Fisher and Tornqvist indices closely approximate each other, but advocates using the Fisher index because it satisfies more desirable axioms for index numbers than does the Tornqvist index. Dumacan (2002) verifies that the Fisher and Tornqvist indices approximate each other when changes in quantities are small, and advocates using the Tornqvist index because it requires less data. Our simulation results using a known CES technology also show that the two indices yield similar values, but the gap between the Fisher index and the Tornqvist index grows as allocative inefficiency increases, and as the corresponding changes in quantities grow.
Figure 4.1 MRTS and $\sigma$

- MRTS at point A
- MRTS at point B
- Weak substitution between inputs
- Strong substitution between inputs
- Actual price vector

$X_1$  $X_2$
Figure 4.2  AI and $\sigma$
Figure 4.3  AI and Indices under CES
Figure 4.4  AI and Indices under Unknown Technology

- TQI \times AI
- FQI \times AI
CHAPTER 5

CONCLUSION

This dissertation is essentially composed of three essays. The first essay shows how to test the Stole-Zwiebel hypothesis that has not been tested since it was proposed in 1996. The second essay shows how to estimate a profit system with technical and allocative inefficiencies by which profit efficiency as an overall indicator of a firm’s performance, is calculated and decomposed to identify the source of efficiency. The third essay addresses unexploited questions about the relation between allocative inefficiency, factor substitutability, and the choice of a quantity index which provides important policy implications. More detailed summaries of the essays follow.

In chapter 2, we present a test of the Stole-Zwiebel hypothesis that individual employees have bargaining power in negotiating compensation and that firms respond by overemploying labor and by adopting inefficient technologies. In the absence of a priori knowledge about the wage bargaining process, a test of the Stole-Zwiebel hypothesis must be inferential. However, their model generates a production function that includes the market wage, rendering problematic the use of standard duality-based estimation procedures to test their predictions of allocative and technical inefficiencies. In this paper, we develop and implement a two-step procedure to test their predictions: the first step determines whether there are allocative inefficiencies in the use of input factors, and the second determines whether any extant allocative inefficiency follows the pattern implied by the Stole-Zwiebel hypothesis. The test is
implemented using data on Korean savings banks. The results indicate the presence of allocative inefficiencies consistent with the Stole-Zwiebel model.

In chapter 3, we propose a way of estimating a profit system with technical and allocative inefficiencies in order to identify the source of profit inefficiency. When technical and allocative inefficiencies are incorporated into the profit system, the existence of unobserved technical inefficiency interactive in a nonlinear form has hindered panel data estimation of the system. To overcome this drawback, homogeneity in technology is introduced to separate technical inefficiency from the profit frontier. Introducing homogeneity in technologies with multiple outputs necessitates choosing a numeraire output, since estimation with homogeneity utilizes a normalized profit function. In our application to Korean savings banks, two numeraires (loans and non-financial assets) are employed to yield different profit efficiencies that are decomposed into technical, allocative and scale inefficiencies and productivity change. A rule of thumb based on careful investigation of the data leads to the conclusion that using non-financial assets as the numeraire provides us with reasonable measures.

In chapter 4, we addressed two questions regarding the relation between allocative inefficiency, substitutability between inputs, and the choice of a quantity index. The first question is how the degree of substitutability between input factors affects the extent of allocative inefficiency and the second question is how well the degree of allocative inefficiency is captured by the Fisher and Tornqvist indices. We showed that allocative inefficiency increases with input substitutability between input factors and that both the Fisher and the Tornqvist indices increase with allocative inefficiency. However, we conclude that the Fisher index captures increased allocative inefficiency more accurately than does the Tornqvist index.


APPENDIX

A. Derivation of the Mean and the Covariance of $\hat{\delta}$

In this appendix we derive the mean and covariance for the estimated vector $\hat{\delta}$. Let the
employment equation (14) be rewritten, for simplicity, as

$$y_{jt} = x_{jt}^T \beta + \alpha_j + \varepsilon_{jt}$$  \hspace{1cm} (A-1)

where $y_{jt} = \log \pi_{jt}$, $x_{jt}^T = [\text{Emp}_{jt}, \text{Emp}_{jt}^2, \text{TE}_j]$ and $\beta^T = [\beta_1, \beta_2, \beta_3]$. The fixed effect data
transformation with $M_j = I - e(e^T e)^{-1} e^T$, where (since we have observations on four time
periods in our sample) $I$ is the 4 x 4 identity matrix and $e$ is a vector of four ones, yields

$$y_j^0 = (x_j^0)^T \beta + \varepsilon_j^0$$  \hspace{1cm} (A-2)

where $y_j^0 = M_j y_{jt}$, $x_j^0 = M_j x_{jt}$, $\varepsilon_j^0 = M_j \varepsilon_{jt}$.

With (A1-2), we can construct the time-demeaning neoclassical profit of firm $j$ at the current
level of employment $n_j$ and the average level of technical efficiency denoted $T\bar{E}$ as

$$y_j^0(n_j, T\bar{E}) = (x_j^0)^T \hat{\beta} + \hat{\varepsilon}_j^0.$$  \hspace{1cm} (A-3)

Under the assumption of fixed effect panel estimation, i.e. $\varepsilon_j^0 \sim N(0, \sigma_{\varepsilon}^2)$ for all $j$, the mean and
the covariance of $y_j^0(n_j, T\bar{E})$ are calculated as

$$E\left(\hat{y}_j^0(n_j, T\bar{E})\right) = (x_j^0)^T \beta \text{ for } \text{Emp}_j = n_j$$
$$\text{Cov}\left(\hat{y}_j^0(n_j, T\bar{E})\right) = \sigma_{\varepsilon}^2 M_j.$$  \hspace{1cm} (A-4)
The mean and covariance of \( \hat{y}_j^0(n_j, T \overline{E}) \), the time-demeaning Stole-Zwiebel profit of firm \( j \) at the current level of employment \( n_j \) and the average level of technical efficiency \( T \overline{E} \), are calculated as

\[
E\left( \frac{\sum_{i=1}^{n_j} y_j^0(i, T \overline{E})}{n_j} \right) = \frac{\sum_{i=1}^{n_j} y_j^0(i, T \overline{E})}{n_j} \sum_{i=1}^{n_j} (x_j^0)^T \beta
\]

\[
\text{Cov}\left( \frac{\sum_{i=1}^{n_j} y_j^0(i, T \overline{E})}{n_j} \right) = \frac{\sigma^2}{n_j} M_j.
\]

(A-5)

Therefore, the logarithm of the firm \( j \)'s time-demeaning front-load factor \( \delta_j^0 \) at the current level of employment \( n_j \) and the average level of technical inefficiency \( T \overline{E} \) has the following mean and covariance,

\[
E(\delta_j^0) = E\left( y_j^0(n_j, T \overline{E}) - \hat{y}_j^0(n_j, T \overline{E}) \right) = (x_j^0)^T \beta - \frac{\sum_{i=1}^{n_j} (x_j^0)^T \beta}{n_j}
\]

\[
\text{Cov}(\delta_j^0) = \text{Cov}\left( y_j^0(n_j, T \overline{E}) - \hat{y}_j^0(n_j, T \overline{E}) \right) = \left( \frac{n_j-1}{n_j} \right) \sigma^2 M_j.
\]

(A-6)

Assuming that one firm’s behavior is independent of the other firms’, the mean and covariance of \( \delta_j^0 \) are

\[
E(\tilde{\delta}_j^0) = E(\hat{\delta}_j^0) = \left[ \begin{array}{c} \tilde{\delta}_j^0 \\ \vdots \\ \tilde{\delta}_N^0 \end{array} \right] \text{ and}
\]
Then, the logarithm of firm j’s estimated front-load factor $\hat{\delta}_j$ is recovered by multiplying $\hat{\delta}_j$ by $M_j$, as are the mean and covariance of $\hat{\delta}_j$.

B. Derivation of the Test Statistic

In this appendix we derive the distribution of test statistic $T S_i$ at (17). We assume that the distribution of $\hat{\delta}$ is normal. Then, under the assumption that $\sigma_\varepsilon^2$ is known, the test statistic $T S_i$ follows a non-central $\chi^2$ with rank of NT-T, where T=4 is the number of time periods in our sample, and non-centrality parameter $\lambda = \frac{1}{2} \delta^T \Sigma^{-1} \delta$:

$$TS_i = \delta^T \Sigma^{-1} \delta \sim \text{non-central } \chi^2_{r, \lambda}.$$  \hfill (B-1)

However, as $\sigma_\varepsilon^2$ is unknown, we use the estimate called for in the fixed effect panel estimation,

$$\hat{\delta}_\varepsilon^2 = \frac{\left( \hat{\varepsilon}_{j\mu} \right)^T \hat{\varepsilon}_{j\mu}}{NT - N - K},$$  \hfill where K=3 and $\hat{\varepsilon}_{j\mu}$ is a vector of residuals obtained from (A-2). Hence,
\[
\frac{\hat{\sigma}^2 (NT - N - K)}{\sigma^2_e} \sim \chi^2_{NT - N - K}.
\] (B-2)

Using (B-2), the test statistic \( TS_1 \) can be expressed as the ratio of two \( \chi^2 \) distributions;

\[
\frac{TS_1}{NT - T} = \left( \frac{\delta^T \Omega^{-1} \delta}{\hat{\sigma}^2_e} \right) / (NT - T)
\]

\[
= \frac{\left( \frac{\delta^T \Omega^{-1} \delta}{\sigma^2_e} \right) / (NT - T)}{\left( \frac{\hat{\sigma}^2_e (NT - N - K)}{\sigma^2_e} \right) / (NT - N - K)}.
\] (B-3)

Finally, since the ratio of \( \chi^2 \) distributions follows an F distribution, (B-3) follows a central F distribution with degrees of (NT-T) and (NT-N-K) under the null hypothesis \( H_0 \) stated at (16),

\[
\frac{TS_1}{NT - T} \sim \text{central } F(NT - T, NT - N - K).
\] (B-4)

C. Proofs of Propositions in Chapter 2

In this appendix we provide proofs of Propositions 1-3 in Chapter 2.

Proof of Proposition 1

The production function \( F(v, z) \) is homogeneous of degree \( r \) in \( v \) if and only if \( F(tv, z) = t^r F(v, z) \) for all \( t > 0 \). Then, the normalized variable profit function can be written as
\[
\hat{\pi}(tw, z) \equiv \max_v \left\{ F(v, z) - \sum_{k=1}^K (tw_k) v_k \right\} \\
= t^{-r} \max_v \left\{ F(tv, z) - \sum_{k=1}^K (t^r w_k) (tv_k) \right\} \\
= t^{-r} \hat{\pi}(t^r w, z).
\]

Hence, we have \( t^r \hat{\pi}(tw, z) = \hat{\pi}(t^{-1} tw, z) \), and by letting \( t^{-1} = \kappa \), so that \( t = \kappa^{r-1} \), we arrive at

\[
\kappa^{r-1} \hat{\pi}(tw, z) = \hat{\pi}(\kappa tw, z),
\]

showing that \( \hat{\pi}(w, z) \) is homogeneous of degree \( \frac{p}{r-1} \) in \( w \).

Proof of Proposition 2

The variable profit function (18) is related to the normalized variable profit function as

\[
\pi(pe^e, w, z) = pe^e \hat{\pi} \left( \frac{w}{pe^e}, z \right).
\]

By letting \( t = \left( pe^e \right)^{-1} \), Proposition 1 allows us to write \( \hat{\pi}(tw, z) = t^{-1} \hat{\pi}(w, z) \) if and only if \( F(v, z) \) is homogeneous of degree \( r \) in \( v \). Thus we have

\[
\pi(pe^e, w, z) = t^{-1} \hat{\pi} \left( tw, z \right) \\
= t^{-\frac{1}{r-1}} \hat{\pi} \left( w, z \right) \\
= \left( pe^e \right)^{\frac{1}{r-1}} \hat{\pi} \left( w, z \right).
\]
Proof of Proposition 3

Linear homogeneity of the variable profit function (18) in the prices \((pe^u, w)\) is expressed as 
\[ \pi(\kappa pe^u, \kappa w, z) = \kappa \pi(pe^u, w, z) \]
for all \(\kappa > 0\). Proposition 2 allows us to rewrite the left-hand side as 
\[ \pi(\kappa pe^u, \kappa w, z) = (\kappa pe^u)\frac{1}{1-r} \hat{\pi}(\kappa w, z) \]
and the right-hand side as 
\[ \kappa \pi(pe^u, w, z) = \kappa \left( pe^u \right)^{\frac{1}{1-r}} \hat{\pi}(w, z) \]
if and only if \(F(v, z)\) is homogeneous of degree \(r\) in \(v\). Thus, we have

\[ (\kappa pe^u)^{\frac{1}{1-r}} \hat{\pi}(\kappa w, z) = \kappa \left( pe^u \right)^{\frac{1}{1-r}} \hat{\pi}(w, z) \]
\[ \hat{\pi}(\kappa w, z) = \kappa^{\frac{r}{r-1}} \hat{\pi}(w, z). \]

Therefore, linear homogeneity of \(\pi(pe^u, w)\) in prices is equivalent to \(\hat{\pi}(w, z)\) being homogeneous of degree \(r/(r-1)\) in \(w\).

D. Proofs of Propositions in Chapter 3

In this appendix we provide proofs of Propositions 1-2 in Chapter 3.

Proof of Proposition 1

Linear homogeneity of the profit function implies
\[
\pi\left(pe^w, w\right) = \left(p_N e^w\right) \pi \left(\hat{p}, \frac{w}{p_N e^w}\right) \\
= \left(p_N e^w\right) \hat{\pi} \left(\hat{p}, \frac{w}{p_N e^w}\right),
\]

\[(D-1)\]

where the second equality follows from the definition of the normalized profit function. If and only if the technology is almost homogeneous of degree \(r,1\) in \((y,x)\), the normalized profit function is homogeneous of degree \(\frac{r}{r-1}\) in \(w\), and (A.1) can be written as

\[
\pi\left(pe^w, w\right) = \left(p_N e^w\right) \left(\frac{1}{p_N e^w}\right)^{\frac{r}{r-1}} \hat{\pi} \left(\hat{p}, w\right)
\]

\[(D-2)\]

Proof of Proposition 2

Linear homogeneity of the profit frontier in \((pe^w, w)\) is expressed as \(\pi(kpe^w, kw) = k\pi(pe^w, w)\). Proposition 1 allows us to rewrite the left-hand side as

\[
\pi(kpe^w, kw) = \left(kp_N e^w\right)^{\frac{1}{r}} \hat{\pi}(\hat{p}, kw) \quad \text{and the right-hand side as} \quad k\pi(pe^w, w) = k \left(p_N e^w\right)^{\frac{1}{r}} \hat{\pi}(\hat{p}, w)
\]

if and only if \(T(y,x) = 0\) is almost homogenous of degree \(r,1\) in \((y,x)\). Thus, we have

\[
\left(kp_N e^w\right)^{\frac{1}{r}} \hat{\pi}(\hat{p}, kw) = k \left(p_N e^w\right)^{\frac{1}{r}} \hat{\pi}(\hat{p}, w)
\]

\[(D-3)\]

\[
\hat{\pi}(\hat{p}, kw) = k^{\frac{r}{r-1}} \hat{\pi}(\hat{p}, w).
\]

Therefore, linear homogeneity of \(\pi(pe^w, w)\) in prices is equivalent to \(\hat{\pi}(\hat{p}, w)\) being homogenous of degree \(\frac{r}{r-1}\) in \(w\).
E. Proofs of the Lemmas and the Proposition in Chapter 4

In this appendix we provide proofs of the lemmas and the Proposition in Chapter 4.

Proof of Lemma 1

Without loss of generality, consider a two-factor, linear homogenous production function in which the perceived (shadow) price of factor two, \( \hat{w}_2 \), exceeds its actual price, \( w \). Then, given any output \( y \), factor demands are \( \hat{x}_i = x_i(\hat{w}, y) \) for \( i = 1, 2 \), and actual cost is \( C^A(\hat{w}) = \hat{x}_1 + w_2 \hat{x}_2 \), where we have normalized the price of factor one to unity. The effect of increased distortion on actual cost is given by

\[
\frac{\partial C^A(\hat{w})}{\partial \hat{w}} = \frac{\partial \hat{x}_1}{\partial \hat{w}} + w_2 \frac{\partial \hat{x}_2}{\partial \hat{w}}
\]

(E-1)

since \( \varepsilon_{i2} = \frac{\partial \hat{x}_i}{\partial \hat{w}} \frac{\hat{w}}{x_i} \) for \( i = 1, 2 \). From Allen (1938, 372-273) [see also Hamermesh (1993, 24)], these own- and cross-price elasticities are given by \( \varepsilon_{22} = -(1-s)\sigma \) and \( \varepsilon_{12} = (1-s)\sigma \) where \( s = \hat{w} \hat{x}_2 / y \). Using these relations, equation (E-1) can be rewritten as

\[
\frac{\partial AC(\hat{w})}{\partial \hat{w}} = \frac{\hat{x}_1}{\hat{w}}(1-s)\sigma - \frac{\hat{x}_2}{\hat{w}}(1-s)\sigma
\]

(E-2)

\[
= (1-s)\sigma \frac{\hat{x}_1 - w \hat{x}_2}{\hat{w}} > 0,
\]

where the inequality follows from
\[ \frac{\partial}{\partial \hat{w}} [\hat{x}_1 + \hat{w}\hat{x}_2] = \hat{x}_2 + \frac{(1-s)\sigma}{\hat{w}} [\hat{x}_1 - \hat{w}\hat{x}_2] = \hat{x}_2, \]  

(E-3)

by the envelope theorem, implying \( \hat{x}_1 = \hat{w}\hat{x}_2 > w\hat{x}_2 \) since \( \hat{w} > w \) is assumed.

Alternatively, if \( \hat{w} < w \), then the inequality in equation (E-2) is reversed. However, since increased distortion in this case implies \( \hat{w} \) declines, actual cost again increases with increased distortion in the input prices.

Proof of Lemma 2

Consider the CES production function, 
\[ y = k(\alpha_1 x_1^\rho + \alpha_2 x_2^\rho)^{\frac{1}{\rho}} \] where \( \alpha_1 + \alpha_2 = 1 \) and \( \sigma = \frac{1}{1-\rho} \) is the elasticity of substitution between factors. The ratio of factor demands is given by

\[ \frac{x_1}{x_2} = \left( \frac{\alpha_1 w}{\alpha_2} \right)^\sigma \]  

(E-4)

which is implied by the equations for the marginal products [see Hamermesh (1993, 24)]. Note that we have again normalized the price of factor one to unity.
As illustrated in the graph above, along the 45 degree line \( x_1 = x_2 = x \), \( y = kx \) and \( x_1 / x_2 = 1 \), regardless of \( \sigma \). However, above or below the 45 degree line, we have

\[
\frac{\partial \left( \frac{x_1}{x_2} \right)}{\partial \sigma} = \left( \frac{\alpha_i w}{\alpha_2} \right)^\sigma \ln \left( \frac{\alpha_i w}{\alpha_2} \right) > 0. \tag{E-5}
\]

Above the 45 degree line we have

\[
\frac{\partial \hat{x}_1}{\partial \sigma} > 0 > \frac{\partial \hat{x}_2}{\partial \sigma}, \tag{E-6}
\]

while below the 45 degree line the inequalities are reversed.

Since factor demands depend on \( \sigma \), we now write actual cost as \( C^{\prime}(\hat{w}, \sigma) \), and since Lemma 1 applies, we have \( \frac{\partial C^{\prime}(\hat{w}, \sigma)}{\partial \hat{w}} = \frac{(1-s)\sigma}{\hat{w}} \) given by equation (E-2). The effect on this rate of change as \( \sigma \) increases is

\[
\frac{\partial}{\partial \sigma} \left[ \frac{\partial \hat{A}(\hat{w}, \sigma)}{\partial \hat{w}} \right] = \left[ \frac{1-s}{\hat{w}} \right] [\hat{x}_1 - w\hat{x}_2] \\
+ \left[ \frac{1-s}{\hat{w}} \right] \sigma \left[ \frac{\partial \hat{x}_1}{\partial \sigma} - w \frac{\partial \hat{x}_2}{\partial \sigma} \right] + \frac{\hat{w} \frac{\partial \hat{x}_2}{\partial \sigma}}{\hat{w} y}. \tag{E-7}
\]

From Lemma 1 and inequalities (E-6), we conclude that each of the three terms on the right-hand side is positive (negative), and that increased distortion entails an increase (a decrease) in \( \hat{w} \), when \( \hat{x}_1 > (\leq) \hat{x}_2 \). Hence, in either case, increased distortion increases actual cost at a faster rate as \( \sigma \) increases. Finally, for the case in which \( \hat{x}_1 = \hat{x}_2 \), the second and third terms on the right hand side of (E-7) vanish. When \( \hat{w} \) exceeds \( w \), the first term on the right hand side is positive and increased distortion entails an increase in \( \hat{w} \). Conversely, when \( w \) exceeds \( \hat{w} \), the first term is
negative and increased distortion entails a reduction in $\hat{w}$. Hence, in both cases, increased distortion increases actual cost at a faster rate as $\sigma$ increases.

Proof of Proposition 1

We again normalize the price of factor one to unity. Since both actual and least-cost factor demands depend on the elasticity of substitution, we write actual cost as $C^A(\hat{w}, \sigma)$ and minimum cost as $C^M(w, \sigma)$. We wish to show that $AI$ increases with $\sigma$, i.e.,

$$\frac{\partial}{\partial \sigma} \left[ \frac{C^A(\hat{w}, \sigma)}{C^M(w, \sigma)} \right] = \left[ \frac{C^A_\sigma}{C^A} - \frac{C^M_\sigma}{C^M} \right] \frac{C^A}{C^M} > 0,$$  \hspace{1cm} (E-8)

for which it suffices to show that the term within brackets on the right hand side of the equality is positive and subscript indicates partial differentiation. We begin by observing that

$$C^A(\hat{w}, \sigma) - C^M(w, \sigma) = \int^W_{\hat{w}} \left( \frac{\partial C^A(\tau, \sigma)}{\partial \hat{w}} \right) d\tau,$$  \hspace{1cm} (E-9)

which implies

$$\frac{\partial}{\partial \sigma} [C^A(\hat{w}, \sigma) - C^M(w, \sigma)] = \int^W_{\hat{w}} \left( \frac{\partial^2 C^A(\tau, \sigma)}{\partial \hat{w} \partial \sigma} \right) d\tau > 0,$$  \hspace{1cm} (E-10)

where the inequality follows from Lemma 2 showing that the cross-partial derivative in the integrand is positive. Inequality (E-10) implies

$$\frac{C^A_\sigma}{C^A} > \frac{C^M_\sigma}{C^M},$$  \hspace{1cm} (E-11)

while $C^A(\hat{w}, \sigma) > C^M(w, \sigma)$ and $C^M_\sigma \leq 0$ together imply

$$\frac{C^M_\sigma}{C^A} \geq \frac{C^M_\sigma}{C^M}.$$  \hspace{1cm} (E-12)
Inequalities (E-11) and (E-12) now imply inequality (E-8).