Technology, productivity and fixed costs : four essays in applied production analysis

Xi Chen

▶ **To cite this version:**

Xi Chen. Technology, productivity and fixed costs : four essays in applied production analysis. Economics and Finance. Université de Strasbourg, 2013. English. NNT : 2013STRAB010 . tel-00998059

HAL Id: tel-00998059

https://theses.hal.science/tel-00998059

Submitted on 30 May 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE AUGUSTIN COURNOT

BUREAU D'ÉCONOMIE THÉORIQUE ET APPLIQUÉE

THÈSE présentée par :

Xi CHEN

soutenue le : **22 novembre 2013**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : sciences économiques

Technologie, productivité et coûts fixes : quatre essais sur l'analyse de la production appliquée

THÈSE dirigée par :

M. KOEBEL Bertrand	professeur, Université de Strasbourg
M. LAISNEY François	professeur, Université de Strasbourg

RAPPORTEURS :

M. MOHNEN Pierre	professeur, UNU-MERIT
M. CHERCHYE Laurens	professeur, KU Leuven

AUTRES MEMBRES DU JURY :

M. LADOUX Norbert	professeur, Université Toulouse 1 Capitole
M. NGUYEN-VAN Phu	chargé de recherche, CNRS

Université de Strasbourg

Faculté des Sciences Économiques et de Gestion

Bureau d'Économie Théorique et Appliquée

THÈSE

de Doctorat en Sciences Économiques

Technology, Productivity and Fixed Costs: Four Essays in Applied Production Analysis

presented by

Xi Chen

Thesis supervisors:

Bertrand Koebel, Professor at University of Strasbourg

François Laisney, Professor at University of Strasbourg

JURY

Laurens Cherchye, Professor at University of Leuven, *External referee*

Bertrand Koebel, Professor at University of Strasbourg

Norbert Ladoux, Professor at University Toulouse I

François Laisney, Professor at University of Strasbourg

Pierre Mohnen, Professor at University of Maastricht, *External referee*

Phu Nguyen-Van, Research Fellow at CNRS, Strasbourg

2013

Acknowledgement

This thesis could not have been accomplished without the support of many people. First and foremost, I am grateful to my doctoral supervisors Bertrand Koebel and François Laisney for their invaluable support, constant encouragement and patient guidance throughout the preparation of this work.

Numerous individuals contributed to this research. In particular, I am indebted to Phu Nguyen-Van, François Fontaine and Rodolphe Dos Santos Ferreira for their constructive comments. My sincere thanks also go to Frédéric Olland for our joint work. I am grateful to the University of Strasbourg, Bureau d'Economie Théorique et Appliquée and Ecole Doctorale Augustin Cournot for providing me with facilities, excellent research environment and financial support.

Last but not least, I want to express my deep appreciation to my parents and to my wife for their unfailing support.

Contents

List of Tables

List of Figures

Chapter 1

General Introduction

> *"These are tasks which will require much time to complete but we submit that they are necessary if the precise relationships which probably lurk within economic phenomena are to be detected and measured."*
>
> — Charles W. Cobb and Paul H. Douglas, A Theory of Production (1928)

> *"For as long as we are unable to put our arguments into the figures, the voice of our science, although occasionally it may help to dispel gross errors, will never be heard by practical men. They are, by instinct, econometricians all of them, in their distrust of anything not amenable to exact proof."*
>
> — Joseph A. Schumpeter, The Common Sense of Econometrics (1933)

This thesis consists of four essays on applied production analysis, with a focus on *Technology*, *Productivity* and *Fixed costs*.[1] It outlines the limits of some recent empirical strategies for modeling producer behavior and proposes several amendments.

1.1 Motivation

In the classical book, *Value and Capital,* Hicks (1946) presents production theory as the study of relationships between productive factors and produced commodities as well as within productive factors. In order to characterize these relationships, production

[1] The four essays are: Chen (2011, 2012); Chen and Koebel (2013); Chen and Olland (2013).

theory relies on the assumption that firms minimize their production costs and maximize the profit, subject to a specific and changing production technology.

One objective of applied production analysis is to represent producer behavior using mathematical formulations, and to conceptualize it through the notions of input substitutability, returns to scale, and technical change among others. A further objective of production analysis is to measure these concepts using production data and statistical methods.

Historically, applied production analysis is originated by the work of Cobb and Douglas (1928) entitled "A Theory of Production". In their approach, the production technology is represented by a parametric function, which later becomes the famous Cobb-Douglas production function. After a heedful work of data collection, Cobb and Douglas fitted their specification using U.S. manufacturing data for the period of 1899-1922. They "materialized", for the first time, the input-output relationship as:[2]

$$P' = 1.01 L^{3/4} C^{1/4}.$$

By today's standards, both the specification and the statistical method used by Cobb and Douglas (1928) seem restrictive and flawed. In the following decades, a series of major contributions to applied production analysis have been made. Frisch (1935) attempted to measure the possibilities of inputs substitution in the manufacture of chocolate. The seminal paper by Arrow, Chenery, Minhas and Solow (1961) proposed the Constant Elasticity of Substitution (CES) production function. Hotelling (1932), Samuelson (1960) and Shephard (1970) received credit for introducing the dual formulation of production theory. Diewert (1971) and Christensen, Jorgenson and Lau (1973) are responsible for the use of flexible functional forms.

Accompanied by the democratization of data and computing power during the last century, applied production analysis and the theory of production have been significantly improved. For instance, the production representation (functional forms) started with the Cobb-Douglas two factors model with two free parameters, augmented to the Translog cost function with four factors and 21 parameters. The two-points flexible form presented in this thesis (Chapter 4) incorporates 36 parameters. The functional forms became fully flexible in the recent years with the consideration of nonparametric functions.

Despite significant developments in this field, we are still facing numerous theoretical and empirical challenges, see for example Jorgenson (1986), Griliches and Mairesse (1995) and Ackerberg et al. (2007). The main purpose of this thesis is to develop empirical strategies with a special focus on production technology specifications and estimation methods. Chapters 2 to 5 extend the literature on production analysis by

[2]Following notations used by Cobb-Douglas (1928), P' denotes the fitted value of production, L and C are the labor and capital index, respectively.

dealing with the following issues.

Stochastic specification In order to formulate an econometric model for fitting the data with theoretical equations, we need to add a (or several) stochastic component(s) to the deterministic functions. Although some properties of production, cost and profit functions can be derived from production theory, it remains silent on the way how the unobserved disturbances should be included into the model. What are the implication of different stochastic specifications?

Substitution in productive factors The CES production function adds flexibility to the Cobb-Douglas specification by treating the elasticity of substitution as an additional free parameter. Despite the simplicity of the CES function, the existing empirical models yield diverging estimation results on the value of the elasticity of substitution. This thesis investigates whether new estimation techniques bring new insights on this elasticity. If the economy was characterized by a CES production function, what could be the bias of considering a Cobb-Douglas specification?

Stochastic technical changes For many specifications used in empirical studies, technical change is assumed to be either homogeneous (using the time trend) or Hicks-neutral. Is it possible to identify the individual-specific bias of technical changes in a flexible way? And how to estimate the factor-augmenting productivity?

Fixed cost and variable cost Christensen et al. (1973) introduced a "transcendental" specification of technology, known as the Translog function, which is able to approximate an arbitrary cost function.[3] The Translog function is considered as a flexible functional form and is still used in many empirical studies. However, an underlying assumption behind this class of flexible function forms is that there is no fixed cost in the long run. In contrast, fixed costs play an important role in numerous theoretical models. Thus, we ask in this thesis: is the Translog cost function adequate for modeling the fixed cost? What are the alternatives?

Applications to other fields Production theory derives fundamental principles which often have profound implications in other fields of economics, in particular for international trade. On the producer side of trade models, productivity and fixed costs of entry play crucial roles for characterizing the equilibrium, see Melitz (2003). Many empirical works studied separately the impacts of productivity and entry costs on firm's trade behavior. This thesis studies the relationship between productivity and fixed costs of entry and their joint effects on the equilibrium. Are more productive firms able to

[3]Christensen et al (1973) proposed the Translog functional form primarily for modeling the production function, whereas this specification is now widely used for the dual cost function.

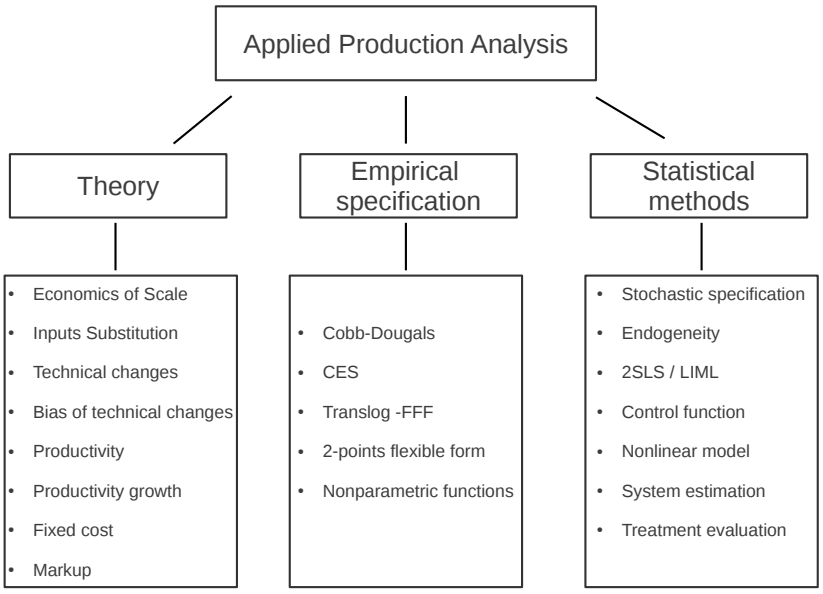


Figure 1.1: An overview of topics studied in this thesis

enter the international market with less entry costs? What are the implication on the equilibrium entry condition?

1.2 An overview of the thesis

Each chapter of this thesis consists of the three building blocks represented in Figure 1.1: the economic theory, the empirical formulation and the statistical methods. Figure 1.1 also provides an overview of different topics studied within each of the three building blocks. Chapters 2 to 4 revisit these building blocks and contribute to one or several of them. Chapter 5 synthesizes the findings and applies them to trade models. In the conclusion chapter, I discuss the limitations of the present work and some potential extensions for further research.

An effort has been made to unify (as much as possible) the theory with the mathematical formulation and the empirical investigation. It reflects my willingness to approach the earlier definition of econometrics by Ragnar Frisch (1933): "[...] econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory [...] Nor should econometrics be taken as synonymous with the application of mathematics to economics. [...] It is the unification of all three that is powerful. And it is this unification that constitutes econometrics". The following paragraphs provides the abstracts of Chapters 2 to 5.

In **Chapter 2**, I conducted a comparative study of four estimators: the OLS, 2SLS, LIML (Limited Information Maximum Likelihood) and F-LIML (Fuller-modified LIML)

estimators. The different estimators are compared for the direct and reverse regression models in the context of estimating returns to scale and technical progress. The chapter shows that the 2SLS estimation method may result in contradictory conclusions when instruments are weak. On the other hand, the LIML and F-LIML estimators provide more reliable estimates. The estimation results based on U.S. manufacturing industry data show that most industries exhibit increasing returns to scale, and the role of technical progress is relatively small when it comes to explaining productivity growth.

Chapter 3 addresses an important issue of empirical production analysis: the estimation of productivity. I extend the Olley-Pakes (1996) estimation method to the CES production function with biased technical change. The new semi-parametric approach allows consistent estimation of the degree of returns to scale, the elasticity of substitution, and the bias in technical change. In this work, I deal with two sources of endogeneity through Hicks-neutral and non-neutral productivity.[4] The proposed method has the advantage that it yields estimates of relative factor-augmenting productivity. The empirical investigation shows strong evidence that U.S. manufacturing industries are characterized by a production technology with elasticity of substitution below one and with a significant bias toward labor-augmenting technical progress.

Chapter 4 investigates the distinction between fixed and variable cost, which is fundamental but quite neglected in production analysis. Empirically, most specifications of production and cost functions assume that fixed costs are zero. This chapter derives the structure of the production function which is necessary and sufficient for generating a fixed cost. We extend the classical production function in order to allow each input to have a fixed and a variable part. Using U.S.manufacturing industry data, we characterize and estimate both fixed and variable components of the cost function. The estimates are used to study how fixed and variable cost interact and affect firms' behavior in terms of price setting and returns to scale.

The aim of **Chapter 5** is to extend the basic Melitz (2003) model by endogenizing the costs of entry into export markets. Our theoretical model of trade behavior highlights the role of heterogeneity in the fixed costs structure and the links between productivity and entry costs. We point out that the Melitz model with homogeneous entry costs may mis-predict the self-selection mechanism. In response to these theoretical findings, we develop an empirical strategy based on the treatment evaluation techniques for measuring costs to entry in the export market at the firm-level. By using French data, our empirical investigation sheds light on the determinants of entry barriers and how they can be reduced from the firm perspective.

[4]In the context of estimating a production function, the endogeneity problem arises because of the correlation of input demands and productivity.

Chapter 2

Direct and Reverse Estimates of Returns to Scale[1]

2.1 Introduction

Increasing returns to scale are of great importance for numerous macroeconomic models (see for example, Farmer and Guo, 1994, and Jones, 2005). However, there is a lack of consensus on whether the assumption of increasing returns to scale is empirically plausible. Different methods of estimating returns to scale have been used in the literature, and produced diverging results. Therefore, it is important to understand exactly what each method does and when it might be preferable to use one over others.

This chapter contributes to the existing literature in several ways. First, I conduct a comparative study between the direct and reverse regression models for various estimators, namely, the Ordinary Least Squares (OLS), the Tow-Stage Least Squares (2SLS), the Limited-Information Maximum Likelihood (LIML) and the Fuller-modified LIML (F-LIML) in the context of estimating returns to scale and technical progress. Second, I point out that the weak instrument problem is an important source of bias, which causes divergence in the estimated value of returns to scale. Our empirical results are produced by using estimation methods that are robust to weak instruments.

An important part of the literature on the estimation of returns to scale relates the output growth index linearly to the input growth index, see for example Diewert and Fox (2008). The intercept and the slope of the linear equation appear as the measurement of technical progress and of returns to scale, respectively. Whereas the theory provides a deterministic relationship between both variables of interest (output and input growth rates), from an empirical perspective it is necessary to decide which variable, the input or the output, is stochastic and therefore measured with errors from the true population regression line. Suppose that the data are represented in a coordinate system, where the

[1]This chapter has been circulated under the title "Increasing returns to scale in U.S. manufacturing industries: evidence from direct and reverse regressions", Chen (2011).

x-axis is the input variable and the y-axis is the output variables. The *direct regression* model assumes that the output variable is stochastic and fits a line that minimizes the squared vertical distance between the data and the regression line in the direction of y-axis. By contrast, the *reverse regression* model considers the reversed situation, where the input variable becomes the dependent variable. Both regression models have theoretical foundation and have been applied in empirical works. Depending on which regression model is chosen, the estimating results are often very different.

Apart from the lack of consensus on regression models, researchers are also debating the choice of estimators. Two types of estimator are considered in the literature, namely, the OLS and Instrumental Variables (IV) estimators, especially the 2SLS estimator. It is easy to show that the direct and reverse OLS estimators are biased when both the dependent variable and the regressor are measured with errors. In contrast, I show in this chapter the direct 2SLS estimator and its reverse counterpart are consistent and asymptotically equivalent. However, in practice the gap between the direct and reverse 2SLS regression is often very large. For instance, Hall (1988) presented the 2SLS estimation results of returns to scale and price-cost markup coefficients using annual two-digit sectoral data for 1953-1984. His estimated returns to scale are unreasonably large for the reverse regression model and even negative for the direct regression model. Bartelsman (1995) was one of the first authors to question the 2SLS estimator used by Hall (1988). Bartelsman provided a series of Monte Carlo experiments to illustrate that the bias is likely to be large when estimating coefficients from the reverse 2SLS approach. An influential article by Basu and Fernald (1997) compared the OLS and 2SLS estimation strategies for the direct regression model. Their OLS estimation results for thirty-four U.S. private business industries (1959-1989) show that estimated coefficients are often much smaller than one (decreasing returns). Returns to scale are larger in the 2SLS estimation, but their average value is still close to one, and cannot confirm the increasing returns to scale hypothesis. By applying the OLS estimation to a larger database (for 1949-2000) of two-digit U.S. manufacturing industries, Diewert and Fox (2008) also obtained contradictory results between the direct and reverse approaches.

One of the reasons for the failure of 2SLS in estimating returns to scale is the weakness of instruments. Studies by Staiger and Stock (1997), Shea (1997) show that weak instruments could generate large bias in the 2SLS estimation. The implausible results of Hall (1988) is a typical consequence of using the weak instruments. Now, the formal diagnostic and cure for weak instruments are available in the literature, see for example, Hahn and Hausman (2002), Stock et al. (2002) and Stock and Yogo (2002). In this chapter, I extend the work of Bartelsman (1995) by taking weak instruments into account. Compared to Hall (1988), Basu and Fernald (1997) and Diewert and Fox (2008), this chapter goes a step further by testing the quality of instruments and estimating the returns to scale with robust estimators, such as the LIML and F-LIML

estimators.

Working with data for twenty one sectors of U.S. manufacturing industries over the last half-century, I found strong evidence of increasing returns to scale. On the other hand, technical progress has made little contribution to U.S. economic growth. Compared with prior empirical results on the estimation of returns to scale, our results are more robust to weak instruments and support a growing body of theoretical models emphasizing the importance of increasing returns to scale for explaining productivity growth.

The remainder of this chapter is organized as follows: I first present the econometric model and the identification issues in Section 2.2. More attention is given to the discussion of IV estimations and the weak instrument problem in Section 2.3. The empirical application to the U.S. manufacturing industries data is reported in Section 2.4. Section 2.5 concludes.

2.2 Econometric model

Based on the prior works of Diewert (1976) and Diewert and Fox (2008), this chapter follows the Diewert-Fox method of measuring technical progress and returns to scale, where a (multiple inputs and multiple outputs) firm's technology is represented by a non-constant returns to scale Translog cost function. The framework proposed by Diewert and Fox (2008) not only relaxes a series of simplifying restrictions, i.e., single-output, constant returns to scale and perfect competition, but also establishes a very practical relationship between aggregate input index and aggregate output index. The measurement of technical progress and of returns to scale appear respectively in this equation as the intercept and the slope, which seem easy to identify. However, I will demonstrate that the identification issue is not straightforward, after a broadly acceptable stochastic specification is chosen. Under a series of restrictions on the Translog cost function and the neutral technical change assumption, the deterministic relationship between the log-Törnqvist input growth index and the log-Törnqvist output growth index is:

$$y^* = \alpha + \beta x^* \,, \tag{2.1}$$

where y^* and x^* are $T \times 1$ vectors of latent values of output growth index and input growth index, respectively. The parameter α is the constant rate of cost reduction and β is the degree of returns to scale. When inputs are increased, if output increases at the same rate, i.e., $\beta = 1$, then the technology exhibits constant returns to scale. If output increases by less than that quantity, i.e., $\beta < 1$, the technology exhibits decreasing returns to scale. If output increases by more than that quantity, i.e., $\beta > 1$, the technology exhibits increasing returns to scale.

The intercept and the slope of equation (2.1) are the two parameters of interest.

Since these factors can never be measured or observed perfectly in the real world, the common practice is to introduce additive error terms. Suppose that there are T observations in the sample, where the observed values are denoted by (x, y). They are measured with additive errors, u and v. Let

$$x = x^* + u \text{ and } \quad y = y^* + v \,, \tag{2.2}$$

where y and x are $T \times 1$ vectors of observations. The model (2.1)-(2.2) is a linear Error-in-Variable (EIV) model, which can be also rewritten in a more compact form with only the observable variables, $y = \alpha + \beta x + \varepsilon$, where $\varepsilon \equiv v - \beta u$. We make some statistical assumptions to restrict our stochastic framework.

Assumptions 2.1. *Suppose u and v are two zero-mean i.i.d. normally distributed variables. Formally, let*

$$\mathrm{V}[u] \equiv \sigma_u^2 \quad \text{and} \quad \mathrm{V}[v] \equiv \sigma_v^2$$

and $V[\varepsilon] = E[(v - \beta u)^2] \equiv \sigma_\varepsilon^2$. The latent variables (x^*, y^*) are uncorrelated with error terms; suppose that the first and second moments exist. Let

$$\mathrm{E}[x^*] \equiv \mu \quad \text{and} \quad \mathrm{V}[x^*] \equiv \sigma^2 \,.$$

The set of parameters that we want to estimate in this model is $\theta \equiv (\alpha, \beta, \mu, \sigma, \sigma_u, \sigma_v)$.

The symmetric treatment of x and y seems to be a simple extension of classical stochastic specification, where only one variable is assumed to be subject to error. The introduction of the second error term increases dramatically the difficulty of estimation. A surprising consequence is that the unique intercept and slope of the fitting line cannot be identified from the bivariate data set (x, y) alone. This is the well-known identification problem of the EIV model, which was firstly highlighted by Adcock (1878) who tried to handle it by using the *Orthogonal regression* (this estimation method is consistent only if both variables are subject to errors that have the same variance, i.e., $\sigma_u^2 = \sigma_v^2$). Adcock's intuition is the origin of *Total least squares* (TLS) estimation, which was generalized one hundred years later by Golub and Van Loan (1980). Another idea on the estimation of measurement error models was introduced by Wald (1940) with the objective of proposing a method in which strong assumptions regarding the error structure are not required. Unfortunately, this class of estimators is not feasible. Since the publication of Wald's method, the problem of estimating EIV models, has received increasing attention from researchers. There have been several surveys, including Madansky (1959), Stefanski (2000) and Gillard (2006).

Under Assumptions 2.1, we can write the following five first and second order moment equations by using the Law of Large Numbers, see Kendall and Stuart (1973):

$$\operatorname{plim} \bar{x} - \mu = 0; \tag{2.3}$$

$$\text{plim}\,\bar{y} - \alpha - \beta\mu = 0; \tag{2.4}$$

$$\text{plim}\,s_{xx} - \sigma^2 - \sigma_u^2 = 0; \tag{2.5}$$

$$\text{plim}\,s_{yy} - \beta^2\sigma^2 - \sigma_v^2 = 0; \tag{2.6}$$

$$\text{plim}\,s_{xy} - \beta\sigma^2 = 0. \tag{2.7}$$

The sample moments of x and y are computed as $\bar{x} = T^{-1}\sum_{t=1}^{T} x_t$; $\bar{y} = T^{-1}\sum_{t=1}^{T} y_t$, $s_{xx} = T^{-1}\sum_{t=1}^{T}(x_t - \bar{x})^2$, $s_{yy} = T^{-1}\sum_{t=1}^{T}(y_t - \bar{y})^2$ and $s_{xy} = T^{-1}\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})$.

The identification problem of EIV models is apparent from an examination of the system (2.3)-(2.7), where there are only five moment equations but six unknown parameters. Thus, we have not enough moment conditions to "fix" the fitting line in order to identify θ. Under Assumptions 2.1, the bivariate data set that contains only the input and output growth index does not provide enough information for consistent estimation of returns to scale and technical progress. However, if one could obtain prior information on parameters then the identification problem can be solved. For example, let $\hat{\alpha}$, $\hat{\beta}$, $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\sigma}_v^2$ and $\hat{\sigma}_u^2$ be consistent estimators of α β, μ, σ^2, σ_v^2 and σ_u^2, respectively. If the variance ratio $\lambda \equiv \sigma_v^2/\sigma_u^2$ is known, we obtain from (2.5), (2.6) and (2.7) a quadratic equation of $\hat{\beta}$:

$$\hat{\beta}^2 s_{xy} - \hat{\beta}(s_{yy} - \lambda s_{xx}) - \lambda s_{xy} = 0\,. \tag{2.8}$$

Given the restriction that $\hat{\beta}$ has the same sign as s_{xy}, there is an unique relevant solution for $\hat{\beta}$. This class of estimator based on prior knowledge of the parameters is called by Wansbeek and Meijer (2000) the *Consistent Adjusted Least Squares* estimation method. Adcock's orthogonal regression (or Total least squares) estimation method is a special case that assumes $\lambda = 1$. Further examples are given in Judge et al., (1980, p.509-531). However, these *ad hoc* assumptions are in general untestable in the empirical economics studies. In order to achieve identification, a classical solution is to introduce a set of instruments and construct the additional moment conditions based on the instruments. In the next section, I first discuss the 2SLS estimators and their potential bias in the context of estimating the degree of returns to scale. This analysis explains the reason of the wide variety of results found in the literature. Then the alternative IV estimation methods, namely the LIML and F-LIML estimators are considered. By following Hahn and Hausman (2002), I compare different estimators for both direct and reverse regression models.

2.3 Direct and reverse regressions

Two types of estimator are commonly used in empirical studies. Some authors, such as Hall (1988, 1990) suggested using the 2SLS estimator. Others like Basu and Fernald (1997), Diewert and Fox (2008) emphasized the OLS approach. It is easy to show that

OLS is biased and that the direct and reverse OLS regressions produce upper and lower bounds on the true value of returns to scale. The direct OLS estimator is defined as $\hat{\beta}_{ols} = s_{xy}/s_{xx}$ with:

$$\text{plim}\,(\hat{\beta}_{ols} - \beta\frac{\sigma^2}{\sigma^2+\sigma_u^2}) = 0.$$

Since the ratio $\hat{\sigma}^2/(\hat{\sigma}^2+\hat{\sigma}_u^2)$ (which is called the *reliability ratio* in the literature) is always less than one, the OLS estimate is downward biased w.r.t. the true slope. The reverse OLS estimator is defined as $\hat{\beta}_{rols} = s_{yy}/s_{xy}$ with:

$$\text{plim}\,(\hat{\beta}_{rols} - \frac{\beta^2\sigma^2+\sigma_v^2}{\beta\sigma^2}) = 0$$

and it is upward biased. The direct and reverse estimates in Diewert and Fox (2008) illustrate the bias of OLS estimator. Unlike the OLS estimator, the 2SLS estimator is in principle able to produce the point estimation by introducing additional instruments into the model, I will show in the following subsection that the direct and reverse 2SLS estimators are consistent and asymptotically equivalent. This result should support the use of 2SLS in empirical research. However, the recent literature recommends extreme caution in interpreting 2SLS estimates, because the 2SLS estimation could lead to large finite-sample bias when the instruments are "weak". Thus, in this study I test the strength of instruments and consider some alternative estimators, which have better finite-sample properties than 2SLS.

2.3.1 Instrumental variable estimator

Assume that we have L instruments such that:

$$\tilde{x} = Z\Pi + \eta, \tag{2.9}$$

$$\tilde{y} = Z\Gamma + \xi, \tag{2.10}$$

where $\tilde{x} \equiv x - \bar{x}$, $\tilde{y} \equiv y - \bar{y}$ and Z is a $T \times L$ matrix of instruments. The errors (η, ξ) are assumed to be i.i.d. normally distributed with $\text{V}[\eta] = \sigma_\eta^2$ and $\text{V}[\xi] = \sigma_\xi^2$. The direct 2SLS estimator of β can be written as:

$$\hat{\beta}_{2sls} = (\tilde{x}'P_Z\tilde{x})^{-1}\tilde{x}'P_Z\tilde{y}\,. \tag{2.11}$$

Similarly, the reverse 2SLS estimator is:

$$\hat{\beta}_{r2sls} = (\tilde{y}'P_Z\tilde{x})^{-1}\tilde{y}'P_Z\tilde{y}\,, \tag{2.12}$$

where $P_Z \equiv Z(Z'Z)^{-1}Z'$ is the orthogonal projection on the column space of Z. Before applying 2SLS to the data, I summarize here the asymptotic properties of the direct

and reverse 2SLS estimator for our EIV model. The formal proofs of Propositions 2.1 and 2.2 are given in Appendix.

Assumption 2.2. *The instruments are orthogonal to the error terms, u, v and ε. The rank condition is satisfied, i.e.,* $\text{rank}\mathrm{E}[Z'Z] = L$ *with* $L \geq 1$, *and* $\text{rank}\mathrm{E}[Z'\tilde{x}] = \text{rank}\mathrm{E}[Z'\tilde{y}] = 1$. *I also assume homoskedasticity to simplify the notations, i.e.,* $\mathrm{E}[\varepsilon^2 Z'Z] = \sigma_\varepsilon^2 \mathrm{E}[Z'Z]$.

Proposition 2.1. *Under Assumptions 2.1 and 2.2, the direct 2SLS estimator defined in* (2.11) *is consistent for* β *and asymptotically normally distributed as:*

$$\sqrt{T}(\hat{\beta}_{2sls} - \beta) \sim N(0, \sigma_\varepsilon^2 A^{-1}), \tag{2.13}$$

where $A \equiv \mathrm{E}(\tilde{x}'Z)[\mathrm{E}(Z'Z)]^{-1}\mathrm{E}(Z'\tilde{x})$ *and* $\mathrm{V}[\varepsilon] = \sigma_\varepsilon^2$.

Proposition 2.2. *Under Assumptions 2.1 and 2.2, the reverse 2SLS estimator defined in* (2.12) *is consistent for* β *and asymptotically normally distributed as:*

$$\sqrt{T}(\hat{\beta}_{r2sls} - \beta) \sim N(0, \sigma_\varepsilon^2 A^{-1}). \tag{2.14}$$

A direct corollary to Propositions 2.1 and 2.2 is that $\sqrt{T}(\hat{\beta}_{2sls} - \beta) \xrightarrow{d} \sqrt{T}(\hat{\beta}_{r2sls} - \beta)$. This result complements Bartelsman's (1995, p.61) finding on the relationship between the direct and reverse 2SLS, which are related to the squared correlation between the projections of $\tilde{x}$ and $\tilde{y}$ on the instruments Z (equation (2.15) below). The next corollary states that this correlation converges asymptotically to 1, and implies that the direct and reverse 2SLS estimators are asymptotically equivalent.

Corollary. *The relationship between* $\hat{\beta}_{2sls}$ *and* $\hat{\beta}_{r2sls}$ *is*

$$\hat{\beta}_{2sls} = \hat{\beta}_{r2sls}[(\tilde{y}'P_Z\tilde{y})^{-1}\tilde{y}'P_Z\tilde{x}(\tilde{x}'P_Z\tilde{x})^{-1}\tilde{x}'P_Z\tilde{y}]. \tag{2.15}$$

Asymptotically

$$(\tilde{y}'P_Z\tilde{x})^{-1}\tilde{y}'P_Z\tilde{x}(\tilde{x}'P_Z\tilde{x})^{-1}\tilde{x}'P_Z\tilde{y} \xrightarrow{p} 1. \tag{2.16}$$

Under the conventional first order asymptotic theory, the direct and reverse 2SLS estimates should be very similar, because the corollary shows that they have unitary correlation. However, the two estimates may differ substantially in practice due to the finite sample bias. The size of this bias depends on several feature of the data and of the underlying regression model.

Finite-sample bias of 2SLS The finite-sample bias is only derived for the direct 2SLS estimator, the bias of the reverse estimator can be carried out in the similar way. We can rewrite the direct 2SLS estimates as:

$$\hat{\beta}_{2sls} = \beta + (\tilde{x}' P_Z \tilde{x}) \tilde{x}' P_Z \varepsilon. \tag{2.17}$$

Taking expectations of $\hat{\beta}_{2sls}$, we see that the bias depends on the terms $\mathrm{E}[(\tilde{x}' P_Z \tilde{x})^{-1}]$ and $\mathrm{E}[\tilde{x}' P_Z \varepsilon]$. For any observation t, we have:

$$\begin{aligned}
\mathrm{E}[\tilde{x}_t' Z_t (Z'Z)^{-1} Z_t' \varepsilon_t] &= \mathrm{E}\{\mathrm{E}[\tilde{x}_t' Z_t (Z'Z)^{-1} Z_t' \varepsilon_t \mid Z]\} \\
&= \mathrm{E}\{\mathrm{E}[\eta_t' Z_t (Z'Z)^{-1} Z_t' \varepsilon_t \mid Z]\} + \mathrm{E}\{\mathrm{E}[\Pi' Z_t \varepsilon_t \mid Z]\} \\
&\overset{a}{=} \sigma_{\eta\varepsilon} \frac{1}{T} \sum_{t=1}^{T} Z_t (Z'Z)^{-1} Z_t' \\
&= \frac{L}{T} \sigma_{\eta\varepsilon},
\end{aligned}$$

where $\sigma_{\eta\varepsilon}$ denotes the covariance between η and ε. The second equality follows because $\tilde{x}_t' = \eta_t' + \Pi' Z_t$ and the instruments Z is independent of error term ε. The last equality follows from the fact that the trace of the projection matrix is equal to the rank of the projection matrix, and the rank of P_z is equal to L. Similarly,

$$\begin{aligned}
\mathrm{E}[\tilde{x}_t' Z_t (Z'Z)^{-1} Z_t' \tilde{x}_t] &= \mathrm{E}\{\mathrm{E}[\tilde{x}_t' Z_t (Z'Z)^{-1} Z_t' \tilde{x}_t \mid Z]\} \\
&= \mathrm{E}[\hat{\Pi} Z_t' Z_t \hat{\Pi}] \\
&\overset{a}{=} \frac{1}{T} \sum_{t=1}^{T} (Z_t' \hat{\Pi})^2.
\end{aligned}$$

We notice from above expressions that the finite-sample bias is caused by the correlation of the errors terms. The bias is monotonically increasing in L (the number of instruments) and monotonically decreasing in $\sum_{t=1}^{T} (Z_t' \hat{\Pi})^2$. The concentration parameter, $\sum_{t=1}^{T} (Z_t' \hat{\Pi})^2 / \sigma_\eta^2$, measures the strength of instruments. Hence, for a fixed number of instruments, highly correlated (with regressors) instruments reduce the finite-sample bias of 2SLS, while weak instruments magnify this bias. Given a value of $\sigma_{\eta\varepsilon}$ and σ_η, the 2SLS finite-sample bias is inversely proportional to $\sum_{t=1}^{T} (Z_t' \hat{\Pi})^2 / \sigma_\eta^2 L$, which can be approximated by the F statistic of the first-stage regression. In order to justify this approximation, I note that the infeasible F statistic follows a non-central $\chi^2(1)$ with non-centrality parameter $\sum_{t=1}^{T} (Z_t' \hat{\Pi})^2 / \sigma_\eta^2 L$. The expectation of feasible F statistic, computed using the estimated value of σ_η, is close to the infeasible F statistic when the sample is larger. Thus, $\mathrm{E}[F] \cong \sum_{t=1}^{T} (Z_t' \hat{\Pi})^2 / \sigma_\eta^2 L + 1$. The following sections discuss the diagnostic of weak instruments based on the first-stage F statistic and present the alternative estimation methods to 2SLS for reducing the finite-sample bias.

2.3.2 Robust estimation inference with weak instruments

A simple response to the finite-sample bias problem is to limit the number of instruments used. One can drop, for example, the weakest instruments selected with help of the first-stage regression or combine them. Given a regression model, the literature suggests that there is a class of alternative estimators, which are asymptotically equivalent to 2SLS but have better finite-sample properties, see Davidson and MacKinnon (1993). One of them is the LIML estimator due to Anderson and Rubin (1949). The LIML estimator is rarely used in applied studies, mainly because the additional assumption of the joint normally of error terms is required. The LIML estimator belongs to the family of k-class estimators proposed by Nagar (1959). The direct and reverse k-class estimators are defined as:

$$\hat{\beta}_{k-class} = [\tilde{x}'(I - kM_z)\tilde{x}]^{-1}\tilde{x}'(I - kM_z)\tilde{y}; \tag{2.18}$$

$$\hat{\beta}_{rk-class} = [\tilde{y}'(I - kM_z)\tilde{x}]^{-1}\tilde{y}'(I - kM_z)\tilde{y}, \tag{2.19}$$

where $M_z = I - P_z$. The direct and reverse LIML estimators are computed as (2.18) and (2.19) by setting k equals to $\hat{k}$, which is estimated as the minimum eigenvalue of the matrix:

$$(W'M_zW)^{-1/2}W'W(W'M_zW)^{-1/2},$$

with $W \equiv (\tilde{y}, \tilde{x})$. The full derivation of the LIML estimator is provided in Davidson and MacKinnon (1993). The OLS and 2SLS estimators are also special cases of the k-class estimator for $k = 0$ and $k = 1$, respectively. One can show $\hat{k}$ converges to 1 at a rate faster than $1/\sqrt{T}$, see Schmidt (1976). Therefore, the LIML estimator is asymptotically equivalent to the 2SLS estimator and Proposition 2.1 and 2.2 also applies to the direct and reverse LIML estimators. Moreover, the direct and reverse LIML estimates coincide when the same instruments set is used for the direct and reverse regressions. In fact, Hahn and Hausman (2002) showed that the LIML estimator is the optimal linear combination of the direct Biased-corrected 2SLS (B2SLS) estimator and the reverse B2SLS estimator, where the B2SLS estimator proposed by Donald and Newey (2001) is defined as the k-class estimator with $k = \frac{(L-2)/T}{1-(L-2)/T}$. Bartelsman (1995) found that the direct and reverse 2SLS estimates are equal, if Z contains only one instrument. In the case of several instruments ($L > 1$), the direct and reverse 2SLS estimates differ, but the direct and reverse LIML estimates are still equal as long as the same set of instruments is used.

Beside the LIML and the Donald-Newey's B2SLS estimators, there are many other k-class estimators. The one needs to be mentioned here, is the optimal Fuller (1977) modified LIML (F-LIML), which is defined as a k-class estimator by setting $k = \hat{k} - 1/(T - L)$. Mariano and Sawa (1972) provided the exact distribution of the LIML estimator and showed that the LIML estimator does not have moments of order greater

than or equal to one in the finite-sample, i.e., $\mathrm{E}[\mid \hat{\beta} \mid] = \infty$.[2] Unlike the LIML estimator, the F-LIML provides all finite moments. Davidson and MacKinnon (1993) showed the distribution of F-LIML lies between the distribution of the 2SLS and LIML estimators. Hahn and Hausman (2002) compared the finite-sample performance of the 2SLS, LIML and F-LIML by Monte Carlo simulations. They found that the unbiased F-LIML is the best estimator in term of MSE. Therefore, the F-LIML is also included for our empirical investigation.

2.3.3 Weak instrument test and specification test

A valid instrument set must satisfy two conditions, the exogeneity and the rank condition, which are formally stated in Assumption 2.2. In addition to the two classical requirements, a valid instrument set must also be highly correlated with endogenous regressors. As mentioned previously, instruments that do not have a high degree of explanatory power, magnify the finite-sample bias of 2SLS. Therefore, a careful diagnostics for weak instruments is important before interpreting the estimation results. Several approaches for testing weak instruments are available. These approaches include the test based on the partial R^2 and the F statistic of the first-stage regression, the test based on the pairwise correlations between the endogenous regressors and instruments, and the Hahn and Hausman (2002) specification test. In this study, I follow Stock and Yogo (2002) weak instruments diagnostic, which uses the first-stage F statistic.

Before going into the econometric analysis of testing weak instruments, it is advisable to define formally the weak instrument set. Stock and Yogo (2002) proposed two definitions based on the type of consequences induced by weak instruments. In general, weak instruments can lead to bias in estimator as shown previously, they can also lead to large size distortion in statistical test. Thus, the first definition of weak instruments set is given in terms of the maximum relative bias, where the relative bias is defined as the ratio of IV bias to OLS bias, i.e., $(\hat{\beta}_{IV} - \beta)/(\hat{\beta}_{OLS} - \beta)$. The second definition is based on the maximum size of Wald test for testing the null hypothesis $\hat{\beta} = \beta$, where the size of test is defined as $\Pr[\text{reject } H_0 \mid H_0 \text{ is true}]$. Hence, a set of instrument is weak if the first-stage F statistic is sufficiently small to cause large relative bias or size distortion.

Given the number of instruments and the definition of weak instruments set, Stock and Yogo (2002, Table 2.1-2.4) provided critical values of F statistic for the 2SLS, LIML and F-LIML estimators. For example, when four instruments are used in a regression model with one endogenous regressor, the F statistic must exceed 6.7 to reject the null hypothesis that the relative bias of 2SLS is larger or equal to 20%. The threshold value of F statistic is increased to 10.3 for a more rigorous test with the null hypothesis that

[2]The nonexistence of finite moments of the LIML estimator, however, should not be considered as the unique criterion in comparing the finite-moment performance with 2SLS. Otherwise, one would always reject the use of LIML in favor to 2SLS.

Table 2.1: Critical values of F statistic given four instruments and one regressor

Bias	2SLS	F-LIML	Size	2SLS	LIML
5%	16.85	7.63	10%	24.58	5.44
10%	10.27	6.37	15%	13.96	3.87
20%	6.71	5.38	20%	10.26	3.30
30%	5.34	4.63	25%	8.31	2.98

Source: Stock and Yogo (2002)

the relative bias is larger or equal to 10%. On the other hand, the 2SLS estimator with a F statistic less than 5, will likely produce biased estimation results. Alternatively, from the size distortion perspective, the F statistic must exceed 24.6 to reject the null hypothesis that the actual size of the 2SLS Wald test at 5% significance level can be greater than 10%. The similar critical values are also available for the F-LIML estimator. However, the LIML estimator does not have moments in finite sample and its relative bias is not well defined, the critical values for testing weak instruments are only given in terms of the maximum size distortion. Since the LIML and F-LIML estimators perform better in finite-samples, their critical values are lower, which suggests that the LIML and F-LIML estimators are superior to 2SLS when the instruments are weak. Table 2.1 summaries the critical values for the weak instrument test with $L = 4$ and the complete tables can be found in Stock and Yogo (2002).

Beside the weak instrument diagnostic, I also test the validity of instrument set with the null hypothesis:

$$H_0 : \mathrm{E}[Z_t \cdot (\tilde{y}_t - \tilde{x}_t\beta)] = 0 \tag{2.20}$$

(or $H_0 : \mathrm{E}[Z_t \cdot (\tilde{x}_t - \tilde{y}_t\beta)] = 0$ in the reverse regression) and the exogeneity of regressors with null hypothesis:

$$H_0 : \mathrm{E}[\tilde{x}_t(\tilde{y}_t - \tilde{x}_t\beta)] = 0 \tag{2.21}$$

(or $H_0 : \mathrm{E}[\tilde{y}_t(\tilde{x}_t - \tilde{y}_t\beta)] = 0$ in the reverse regression). The validity of instruments can be tested only in the over-identified models. The rejection of the null hypothesis (2.20) means that at least one of the instruments is not valid. The failure to reject the null hypothesis (2.20) guarantees only the over-identification restrictions are valid.

The test for regressor exogeneity is conditional on the consistency of 2SLS estimation. The ideas is to use a Hausman specification test to see whether there is significant difference between the OLS estimate and the consistent 2SLS estimate. A significant difference implies the rejection of null hypothesis (2.21), which indicates the endogeneity of regressor. Otherwise, the OLS estimate is consistent. However, one should be aware that the 2SLS estimator is biased toward OLS in the situation of weak instruments. Thus, the failure to reject the null hypothesis (2.21) may be because of the instruments Z are weak. For both tests, I consider the robust (HAC) estimator of the covariance matrix. The corresponding statistics are the Hansen's J statistic for testing

over-identification restriction and the Durbin-Wu-Hausman (DWH) statistic for the regressor exogeneity. The Hansen's J statistic has an asymptotic chi-square distribution with degree of freedom equals the number of over-identification restrictions. The DWH statistic follows a chi-square distribution with degree of freedom equals one (the number of regressors).

2.4 Evidence of increasing returns to scale

The data set used in this work comes from the U.S. Bureau of Labor Statistics (BLS), especially the historical KLEMS database for 1949-2001.[3] The BLS's Multifactor Productivity program provides the annual output and combined input quantity indexes, which are calculated using the Törnqvist index formula. The combined input index is an aggregate of capital, labor, energy, material and purchased business services inputs. The associated output and input prices are also provided. Twenty one manufacturing sectors including three aggregate sectors are considered in this study. The input and output growth indexes x and y correspond to the first-difference of inputs and output quantity indexes, respectively. The changes (first-differences) in the price of capital, labor, energy, material and services make up my set of five instruments. The input prices expressed in level have also been considered as instrument, but the first-stage F statistic shows that the prices expressed in first-difference are more relevant.

The direct and reverse regression models are estimated using the OLS, 2SLS, LIML and F-LIML estimators. In order to limit the number of instruments, the weakest instrument is dropped. The main outcomes of these estimations are summarized in Table 2.2, 2.3 and 2.4, which report the estimates of returns to scale and technical progress parameters as well as the estimated standard errors and the test statistics.

The estimates of returns to scale from the direct OLS regression average to 1.1 and the reverse OLS estimates average to 1.4, the average difference between direct and reverse estimates is about 0.3. This results is similar to those found by Diewert and Fox (2008) and Koebel and Laisney (2010) who use the same data. The difference between the direct and reverse estimates is mainly due to the opposite bias of direct and reverse OLS estimators. In some sectors, the two OLS estimators produced very contradictory results. For instance, the direct OLS regression suggests a statistically significant decreasing returns to scale of 0.5 for the "Food & Kindred Prod" sector (SIC 20). On the other hand, the reverse OLS reports an increasing returns to scale of 2.1 for the same sector.

Both OLS estimates are inconsistent in the EIV framework due to the endogenous regressor. The IV estimations (2SLS, LIML and F-LIML) can solve the endogeneity problem as long as the instruments set is not weak. I report in Table 2.2 the first-stage

[3]See BLS website http://www.bls.gov/mfp/. The data for 1950, 1951 and 1952 are missing. An older version of this data set was used by Diewert and Fox (2008).

Table 2.2: OLS and 2SLS estimates of returns to scale

	Direct regression						Reverse regression					
SIC	OLS	2SLS	DWH	Hansen	F	WI	OLS	2SLS	DWH	Hansen	F	WI
Manu.	1.315 ***	1.467 ***	5.580	4.502	11.954	z4	1.456 ***	1.510 ***	1.141	3.573	16.690	z4
	(0.048)	(0.076)	(0.018)	(0.212)			(0.068)	(0.106)	(0.285)	(0.311)		
Nondur.	1.193 ***	1.736 ***	3.441	1.763	3.080	z5	1.659 ***	1.856 ***	1.401	1.397	6.082	z5
	(0.176)	(0.161)	(0.064)	(0.623)			(0.136)	(0.198)	(0.237)	(0.706)		
20	0.468 ***	0.834 **	1.889	2.773	4.683	z3	2.083 ***	1.719 **	0.431	7.269	7.308	z1
	(0.126)	(0.318)	(0.169)	(0.428)			(0.503)	(0.591)	(0.512)	(0.064)		
22	0.951 ***	1.056 ***	3.082	4.157	8.299	z4	1.060 ***	1.070 ***	0.031	4.015	10.230	z4
	(0.045)	(0.084)	(0.079)	(0.245)			(0.056)	(0.085)	(0.861)	(0.260)		
23	0.973 ***	0.880 ***	0.249	7.981	0.616	z3	1.096 ***	2.076	2.601	2.997	1.069	z2
	(0.038)	(0.208)	(0.618)	(0.046)			(0.077)	(1.152)	(0.107)	(0.392)		
26	1.253 ***	1.318 ***	0.332	1.938	13.834	z4	1.669 ***	1.414 ***	2.229	1.851	10.859	z4
	(0.133)	(0.154)	(0.565)	(0.585)			(0.126)	(0.138)	(0.136)	(0.604)		
27	1.224 ***	1.331 ***	0.345	10.617	6.046	z4	1.616 ***	1.763 ***	0.460	3.541	8.000	z4
	(0.118)	(0.295)	(0.557)	(0.014)			(0.158)	(0.387)	(0.497)	(0.315)		
28	1.166 ***	1.498 ***	3.731	10.107	4.780	z3	1.758 ***	2.249 **	3.751	3.237	10.913	z3
	(0.171)	(0.190)	(0.053)	(0.018)			(0.257)	(0.801)	(0.053)	(0.357)		
29	1.210 ***	1.251 ***	0.012	10.281	6.875	z3	1.255 ***	1.319 ***	0.268	7.509	7.893	z3
	(0.031)	(0.082)	(0.914)	(0.016)			(0.055)	(0.136)	(0.604)	(0.057)		
30	1.109 ***	1.296 ***	1.923	2.005	3.830	z2	1.207 ***	1.367 ***	1.749	1.119	5.563	z5
	(0.042)	(0.171)	(0.166)	(0.571)			(0.054)	(0.211)	(0.186)	(0.772)		
Durab.	1.250 ***	1.326 ***	4.432	6.498	17.477	z5	1.322 ***	1.351 ***	0.622	5.264	21.815	z5
	(0.025)	(0.066)	(0.035)	(0.090)			(0.049)	(0.080)	(0.430)	(0.153)		
24	0.827 ***	0.965 ***	3.742	2.287	13.653	z3	1.120 ***	0.983 ***	2.698	2.489	14.385	z5
	(0.064)	(0.084)	(0.053)	(0.515)			(0.080)	(0.082)	(0.101)	(0.477)		
25	1.150 ***	1.305 ***	7.019	8.841	4.393	z2	1.204 ***	1.369 ***	5.424	6.394	6.410	z2
	(0.044)	(0.092)	(0.008)	(0.031)			(0.048)	(0.124)	(0.020)	(0.094)		
32	1.277 ***	1.668 ***	7.174	3.974	5.861	z3	1.452 ***	1.778 ***	8.276	4.607	13.924	z5
	(0.077)	(0.168)	(0.007)	(0.264)			(0.065)	(0.219)	(0.004)	(0.203)		
33	1.215 ***	1.277 ***	1.247	2.561	11.910	z4	1.285 ***	1.296 ***	0.046	4.211	13.490	z5
	(0.041)	(0.059)	(0.264)	(0.464)			(0.036)	(0.063)	(0.831)	(0.240)		
34	1.123 ***	1.213 ***	4.485	2.656	9.929	z4	1.163 ***	1.222 ***	2.580	2.186	12.825	z4
	(0.022)	(0.082)	(0.034)	(0.448)			(0.034)	(0.086)	(0.108)	(0.535)		
35	1.136 ***	1.253 ***	2.043	5.832	5.670	z5	1.265 ***	1.312 ***	0.267	4.830	7.010	z5
	(0.038)	(0.112)	(0.153)	(0.120)			(0.058)	(0.143)	(0.605)	(0.185)		
36	1.195 ***	1.226 ***	0.131	5.035	4.800	z5	1.283 ***	1.297 ***	0.017	5.453	5.053	z5
	(0.041)	(0.113)	(0.718)	(0.169)			(0.077)	(0.169)	(0.898)	(0.141)		
37	1.151 ***	1.185 ***	0.685	11.678	7.331	z4	1.193 ***	1.238 ***	0.666	9.299	8.208	z4
	(0.030)	(0.064)	(0.041)	(0.009)			(0.046)	(0.104)	(0.415)	(0.026)		
38	1.012 ***	1.083 ***	2.424	12.890	16.420	z4	1.113 ***	1.140 ***	0.524	10.332	20.460	z2
	(0.079)	(0.046)	(0.120)	(0.005)			(0.023)	(0.039)	(0.469)	(0.016)		
39	1.050 ***	0.436	4.581	4.655	1.463	z3	1.555 ***	5.362	4.366	2.864	1.752	z5
	(0.090)	(0.688)	(0.032)	(0.199)			(0.194)	(8.995)	(0.037)	(0.413)		
mean	1.107	1.258	-	-	7.757	-	1.372	1.434	-	-	9.997	-

Note: The column "WI" reports the dropped instrument. z1: price of capital; z2: price of labor; z3: price of energy; z4: price of material; z5: price of service. For the estimated coefficient, the standard errors are reported in parenthesis. For the test statistics, the associated p-values are reported in parenthesis. * corresponds to p<0.05, ** corresponds to p<0.01, *** corresponds to p<0.001. The mean values in the last row are computed by averaging over estimates that are at least significant at the 5% threshold.

Table 2.3: LIML and F-LIML estimates of returns to scale

	Direct regression				Reverse regression			
SIC	LIML	F-LIML	Hansen (LIML)	Hansen (F − LIML)	LIML	F-LIML	Hansen (LIML)	Hansen (F − LIML)
Manu.	1.530 ***	1.520 ***	3.291	3.438	1.530 ***	1.527 ***	3.283	3.321
	(0.133)	(0.123)	(0.349)	(0.329)	(0.133)	(0.129)	(0.350)	(0.345)
Nondur.	1.891 ***	1.815 ***	1.294	1.514	1.891 ***	1.875 ***	1.360	1.377
	(0.254)	(0.198)	(0.731)	(0.679)	(0.238)	(0.218)	(0.715)	(0.711)
20	1.124 *	1.062 *	2.958	2.935	1.576 *	1.600 *	6.739	6.832
	(0.513)	(0.467)	(0.398)	(0.402)	(0.734)	(0.709)	(0.081)	(0.077)
22	1.071 ***	1.065 ***	4.421	4.323	1.071 ***	1.071 ***	4.030	4.024
	(0.093)	(0.089)	(0.219)	(0.229)	(0.091)	(0.089)	(0.258)	(0.259)
23	−8.502	0.742	0.037	5.740	−7.890	4.002	0.708	1.586
	(781.252)	(0.672)	(0.998)	(0.125)	(77.415)	(8.308)	(0.871)	(0.663)
26	1.339 ***	1.336 ***	1.903	1.909	1.339 ***	1.351 ***	1.803	1.811
	(0.170)	(0.167)	(0.593)	(0.591)	(0.170)	(0.164)	(0.614)	(0.612)
27	2.572	2.208	2.078	3.205	2.572	2.429	2.438	2.576
	(8.590)	(4.986)	(0.556)	(0.361)	(3.133)	(2.468)	(0.487)	(0.462)
28	3.153	2.749	2.881	3.922	3.153	3.033	2.205	2.277
	(7.501)	(4.723)	(0.410)	(0.270)	(3.313)	(2.878)	(0.531)	(0.517)
29	2.061	1.776	2.405	4.045	2.061	1.877	4.251	4.979
	(10.927)	(5.020)	(0.493)	(0.257)	(6.146)	(3.824)	(0.236)	(0.173)
30	1.377 ***	1.350 ***	1.507	1.652	1.403 ***	1.388 ***	1.120	1.121
	(0.275)	(0.237)	(0.681)	(0.648)	(0.264)	(0.240)	(0.772)	(0.772)
Durab.	1.365 ***	1.361 ***	5.032	5.150	1.365 ***	1.364 ***	4.852	4.889
	(0.105)	(0.101)	(0.169)	(0.161)	(0.101)	(0.099)	(0.183)	(0.180)
24	0.973 ***	0.967 ***	2.286	2.287	0.974 ***	0.979 ***	2.482	2.486
	(0.086)	(0.084)	(0.515)	(0.515)	(0.084)	(0.083)	(0.479)	(0.478)
25	1.582 **	1.525 ***	3.056	3.748	1.582 ***	1.544 ***	3.563	3.887
	(0.566)	(0.434)	(0.383)	(0.290)	(0.443)	(0.372)	(0.313)	(0.274)
32	1.816 ***	1.778 ***	2.846	3.090	1.871 ***	1.850 ***	4.143	4.237
	(0.287)	(0.252)	(0.416)	(0.378)	(0.300)	(0.280)	(0.246)	(0.237)
33	1.291 ***	1.288 ***	2.513	2.525	1.300 ***	1.300 ***	4.233	4.230
	(0.067)	(0.065)	(0.473)	(0.471)	(0.074)	(0.072)	(0.237)	(0.238)
34	1.236 ***	1.231 ***	2.042	2.168	1.236 ***	1.233 ***	1.879	1.944
	(0.110)	(0.103)	(0.564)	(0.538)	(0.105)	(0.101)	(0.598)	(0.584)
35	1.342 ***	1.325 ***	4.645	4.830	1.342 ***	1.336 ***	4.592	4.630
	(0.239)	(0.211)	(0.200)	(0.185)	(0.220)	(1.336)	(0.204)	(0.201)
36	1.345	1.319 *	4.098	4.269	1.345 *	1.335 *	5.260	5.292
	(0.705)	(0.533)	(0.251)	(0.234)	(0.671)	(0.555)	(0.154)	(0.152)
37	1.670	1.538	2.825	4.139	1.670	1.579	4.125	4.728
	(4.460)	(2.568)	(0.419)	(0.247)	(2.974)	(2.044)	(0.248)	(0.193)
38	1.129 ***	1.124 ***	12.143	12.241	1.173 ***	1.171 ***	10.202	10.211
	(0.046)	(0.044)	(0.007)	(0.007)	(0.085)	(0.081)	(0.017)	(0.017)
39	−1.405	−0.432	1.311	2.570	−2.295	−6.136	1.183	1.692
	(8.041)	(3.118)	(0.726)	(0.463)	(7.097)	(29.705)	(0.757)	(0.639)
mean	1.362	1.338	-	-	1.400	1.395	-	-

Note: see Table 2.2.

Table 2.4: Estimates of technical change

	Direct regression				Reverse regression			
SIC	OLS	2SLS.	LIML	F-LIML	OLS	2SLS	LIML	F-LIML
Manu.	0.006 *	0.003	0.002	0.002	0.003	0.002	0.002	0.002
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
Nondur.	0.002	−0.008*	−0.011*	−0.010*	−0.007	−0.011*	−0.011	−0.011
	(0.003)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)
20	0.014 ***	0.007	0.001	0.002	−0.017	−0.010	−0.007	−0.008
	(0.003)	(0.007)	(0.011)	(0.010)	(0.011)	(0.013)	(0.015)	(0.015)
22	0.024 ***	0.024 ***	0.024 ***	0.024 ***	0.024 ***	0.024 ***	0.024 ***	0.024 ***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
23	0.011 ***	0.011 ***	0.078	0.012 *	0.010 ***	0.003	0.074	−0.011
	(0.002)	(0.002)	(5.586)	(0.005)	(0.002)	(0.012)	(0.542)	(0.067)
26	−0.000	−0.002	−0.002	−0.002	−0.011	−0.004	−0.002	−0.003
	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
27	−0.010*	−0.013	−0.047	−0.037	−0.020***	−0.025*	−0.047	−0.043
	(0.004)	(0.009)	(0.237)	(0.137)	(0.005)	(0.010)	(0.085)	(0.066)
28	0.005	−0.006	−0.057	−0.044	−0.014	−0.029	−0.057	−0.053
	(0.006)	(0.006)	(0.225)	(0.140)	(0.011	(0.024)	(0.097)	(0.084)
29	−0.000	−0.001	−0.018	−0.012	−0.001	−0.003	−0.018	−0.014
	(0.002)	(0.002)	(0.221)	(0.101)	(0.002)	(0.002)	(0.123)	(0.076)
30	0.003	−0.004	−0.007	−0.006	−0.001	−0.007	−0.009	−0.008
	(0.004)	(0.008)	(0.012)	(0.010)	(0.005)	(0.010)	(0.011)	(0.011)
Durab.	0.011 ***	0.010 **	0.009 *	0.009 *	0.010 **	0.009 *	0.009 *	0.009 *
	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
24	0.013 **	0.011 *	0.011 *	0.011 *	0.010	0.011 *	0.011 *	0.011 *
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
25	0.003	−0.001	−0.007	−0.006	0.002	−0.002	−0.007	−0.006
	(0.003)	(0.003)	(0.013)	(0.010)	(0.003)	(0.003)	(0.010)	(0.008)
32	0.002	−0.005	−0.007	−0.006	−0.001	−0.006	−0.008	−0.008
	(0.003)	(0.005)	(0.006)	(0.006)	(0.004)	(0.006)	(0.007)	(0.006)
33	−0.001	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002	−0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
34	0.000	−0.002	−0.002	−0.002	−0.001	−0.002	−0.002	−0.002
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
35	0.018 ***	0.014 *	0.011	0.012	0.014 **	0.012	0.011	0.011
	(0.005)	(0.006)	(0.008)	(0.008)	(0.005)	(0.006)	(0.008)	(0.008)
36	0.025 ***	0.024 ***	0.019	0.020	0.022 ***	0.021 **	0.019	0.020
	(0.004)	(0.006)	(0.028)	(0.021)	(0.005)	(0.008)	(0.027)	(0.022)
37	0.005	0.004	−0.009	−0.005	0.004	0.003	−0.009	−0.006
	(0.004)	(0.004)	(0.110)	(0.062)	(0.004)	(0.004)	(0.071)	(0.048)
38	0.015 ***	0.012 *	0.010	0.010	0.011 **	0.009	0.008	0.008
	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.005)	(0.006)	(0.006)
39	0.010	0.018	0.045	0.031	0.002	−0.053	0.058	0.111
	(0.005)	(0.010)	(0.113)	(0.043)	(0.006)	(−0.053)	(0.111)	(0.446)
mean	0.013	0.012	0.008	0.009	0.010	0.005	0.015	0.015

Note: see Table 2.2.

F statistic for both direct and reverse regression models (for each of the two endogenous regressors, $\tilde{x}$ and $\tilde{y}$, I use the same set of instruments). The average first-stage F statistic by dropping the weakest instrument reaches 7.8 for the direct regression and 10.0 for the reverse regression, which are lower than and equal to the threshold of 10. Hence, we should suspect that the 2SLS estimator may generate more than 20% relative bias and 15% size distortion (for a 2SLS Wald test at 5% significance level). However, the weak instrument problem is less severe in the LIML and F-LIML estimation. For both direct and reverse regression models, the F statistic is sufficiently larger that the relative bias of F-LIML estimates is less than 5% and that the size distortion of LIML estimators is less than 5%. In some sectors, such as "Durable Goods" ($F = 17.5$ for the direct regression and $F = 21.8$ for the reverse regression), the F statistic is large enough that the 2SLS estimator is approximately unbiased (less than 5% relative bias) with a small size distortion (less than 10% size distortion). There are other sectors, such as "Apparel and Related prod." (SIC 23) with $F = 0.6$ for the direct regression and $F = 1.1$ for the reverse regression, the F statistic is too low that the test fails to reject the null hypothesis of weak instruments even for the LIML and F-LIML estimators.

Given the results of weak instrument test, now I examine the estimates of returns to scale. The 2SLS estimates are reported in Table 2.2, the LIML and F-LIML estimates are reported in Table 2.3. The 2SLS estimates suggest larger degree of returns to scale than OLS, and all 2SLS estimates of returns to scale are significantly non-decreasing. The 2SLS Hansen's J statistic (for testing the null hypothesis that the instruments set is valid) and the 2SLS Durbin-Wu-Hausman statistic (for testing the null hypothesis that the regressors are exogenous) as well as the associated p-values in parenthesis are also reported in Table 2.2. The direct 2SLS Hansen test rejects the validity of instruments set in seven sectors and the reverse 2SLS Hansen test rejects the null hypothesis in two sectors (at 5% significance level). In the majority of sector the test fails to reject that the OLS estimates are significantly different from the 2SLS estimates. This result is not surprising, because with weak instruments, the 2SLS estimators is biased toward OLS. The direct and reverse 2SLS estimators of β are in general closer than those obtained by OLS, except for two sectors, "Apparel and Related prod." (SIC 23) and "Misc. Manufacturing" (SIC 39), which have low first-stage F statistic. Theoretically, according to Proposition 2.1 and 2.2, the direct and reverse 2SLS are asymptotically equivalent in the conventional sense. Nevertheless, some difference between the two estimates is due to the finite-sample bias, which is magnified when instruments are weak. In the case of "Food and Kindred Prod." sector (SIC 20, $F = 4.7$ for the direct regression and $F = 7.3$ for the reverse regression), the difference between the two estimates is larger than one, where the direct 2SLS regression suggests decreasing returns to scale and the reverse 2SLS regression provides the opposite results. These results are somewhat contradictory but more plausible than these obtained by Hall (1990), who

reported direct estimated returns to scale of 0.3 and reverse estimated returns to scale of 33.5. The evidence from Table 2.2 indicates the presence of weak instrument problem in our data. Hence, we need to use in this case, more robust estimation methods, namely the LIML and F-LIML estimators.

The LIML and F-LIML estimators reduce substantially the gap between the direct regression and the reverse regression. The average direct and reverse difference is 0.04 for LIML and 0.06 for F-LIML, which are about 20% and 30% of those produced by OLS and 2SLS. Hence, we avoid the contradiction between the two regression models. For instance, one can safely say that the "Food and kindred Prod." sector (SIC 20) is characterized by a production technology with non-decreasing returns to scale, which is range between 1.1 and 1.6. Closing the gap between direct and reverse estimates can be viewed as an evidence that the LIML and F-LIML estimators are more robust to weak instruments than 2SLS. Indeed, the weak instrument test proposed by Hahn and Hausman (2002) is based on testing the difference between the direct and reverse estimates.

I report in Table 2.4 the direct and reverse estimates of technical change, $\hat{\alpha}$. Similarly to Diewert and Fox (2008), our estimates of α are low and insignificantly for many sectors, which suggest that the effect of technical progress is modest when it comes to explaining productivity growth. However, there is an exception, the technical change parameter is positive and statistically significant for "Textile Mill prod." sector (SIC 22) whereas returns to scale is not significantly different from one. This results shows that, unlike other sectors, the source of productivity growth for "Textile Mill prod." sector comes from positive technical progress rather than increasing returns to scale.

2.5 Conclusion

Following the theoretical developments, the empirical outcomes of this chapter provide evidence of increasing returns to scale and relatively small technical progress. Compared with prior studies, including Hall (1991), Bartelsman (1995), Basu and Fernald (1997) and Diewert and Fox (2008), this chapter contains a more complete econometric analysis and yield more convincing empirical results for the estimated coefficients of returns to scale.

This chapter compares the OLS, 2SLS, LIML and F-LIML estimators for the direct and reverse regression models within the EIV framework. I show that the reliability of 2SLS estimates is highly dependent on the strength of instruments. Our empirical findings suggest that the LIML and F-LIML are clearly better estimation methods when instruments are weak. The gap between the direct and reverse LIML and F-LIML estimates are generally smaller than those of OLS and 2SLS estimates. The next step of research on this topic may be oriented towards exploring richer databases, such

as the four-digit NAICS manufacturing industries database from BLS or the six-digit NAICS database from NBER-CES, and using panel data analysis. A further purpose is to generalize these findings to nonlinear specifications of the production technology along the lines of Kumbhakar and Tsionas (2011).

2.6 Appendix

Asymptotic Properties of the direct and reverse 2SLS estimators

Proof of Proposition 2.1.

The direct 2SLS estimator of our model can be written as*:*

$$\hat{\beta}_{2sls} = [(\sum_{t=1}^{T} \tilde{x}_t' Z_t)(\sum_{t=1}^{T} Z_t' Z_t)^{-1}(\sum_{t=1}^{T} Z_t' \tilde{x}_t)]^{-1}(\sum_{t=1}^{T} \tilde{x}_t' Z_t)(\sum_{t=1}^{T} Z_t' Z_t)^{-1}(\sum_{t=1}^{T} Z_t' \tilde{y}_t) \, .$$

Since the observable output index is $\tilde{y}_t = \beta \tilde{x}_t + \varepsilon_t$, we can rewrite the estimator as*:*

$$\hat{\beta}_{2sls} = \beta + [(\sum_{t=1}^{T} \tilde{x}_t' Z_t)(\sum_{t=1}^{T} Z_t' Z_t)^{-1}(\sum_{t=1}^{T} Z_t' \tilde{x}_t)]^{-1}(\sum_{t=1}^{T} \tilde{x}_t' Z_t)(\sum_{t=1}^{T} Z_t' Z_t)^{-1}(\sum_{t=1}^{T} Z_t' \varepsilon_t) \, .$$

Under Assumption 2.2 and the law of large numbers,

$$(\sum_{t=1}^{T} \tilde{x}_t' Z_t)(\sum_{t=1}^{T} Z_t' Z_t)^{-1}(\sum_{t=1}^{T} Z_t' \tilde{x}_t)$$

is non singular and we have:

$$\text{plim}[(\sum_{t=1}^{T} \tilde{x}_t' Z_t)(\sum_{t=1}^{T} Z_t' Z_t)^{-1}(\sum_{t=1}^{T} Z_t' \varepsilon_t)] = 0 \, .$$

Thus, we have $\text{plim}[\hat{\beta}_{2sls}] = \beta + A^{-1}0 = \beta$. We can write:

$$\sqrt{T}(\hat{\beta}_{2sls} - \beta) = [(T^{-1}\textstyle\sum_{t=1}^{T} \tilde{x}_t' Z_t)(T^{-1}\sum_{t=1}^{T} Z_t' Z_t)^{-1}(T^{-1}\sum_{t=1}^{T} Z_t' \tilde{x}_t)]^{-1}$$

$$(T^{-1}\textstyle\sum_{t=1}^{T} \tilde{x}_t' Z_t)(T^{-1}\sum_{t=1}^{T} Z_t' Z_t)^{-1}(T^{-1/2}\sum_{t=1}^{T} Z_t' \varepsilon_t) \, .$$

Then,

$$[(T^{-1}\sum_{t=1}^{T} \tilde{x}_t' Z_t)(T^{-1}\sum_{t=1}^{T} Z_t' Z_t)^{-1}(T^{-1}\sum_{t=1}^{T} Z_t' \tilde{x}_t)]^{-1} - A^{-1} = o_p(1) \, .$$

The central limit theorem implies that

$$(T^{-1}\sum_{t=1}^{T} \tilde{x}_t' Z_t)(T^{-1}\sum_{t=1}^{T} Z_t' Z_t)^{-1}(T^{-1/2}\sum_{t=1}^{T} Z_t' \varepsilon_t) \sim N(0, B) \, ,$$

where $B \equiv E(\tilde{x}'Z)E(Z'Z)^{-1}E(\varepsilon^2 Z'Z)E(Z'Z)^{-1}E(Z'\tilde{x})$. Therefore,

$$\sqrt{T}(\hat{\beta}_{2sls} - \beta) = A^{-1}(T^{-1}\sum_{t=1}^{T}\tilde{x}_t'Z_t)(T^{-1}\sum_{t=1}^{T}Z_t'Z_t)^{-1}(T^{-1/2}\sum_{t=1}^{T}Z_t'\varepsilon_t) + o_p(1)$$

and

$$\sqrt{T}(\hat{\beta}_{2sls} - \beta) \sim N(0, A^{-1}BA^{-1}) .$$

The homoskedasticity assumption allows to simplifying the form of 2SLS asymptotic variance to $\sigma_\varepsilon^2 A^{-1}$.

Proof of Proposition 2.2.

The reverse 2SLS estimator is

$$\hat{\beta}_{r2sls} = [(\sum_{t=1}^{T}\tilde{y}_t'Z_t)(\sum_{t=1}^{T}Z_t'Z_t)^{-1}(\sum_{t=1}^{T}Z_t'\tilde{x}_t)]^{-1}(\sum_{t=1}^{T}\tilde{y}_t'Z_t)(\sum_{t=1}^{T}Z_t'Z_t)^{-1}(\sum_{t=1}^{T}Z_t'\tilde{y}_t)$$

$$= \beta + [(\sum_{t=1}^{T}\tilde{y}_t'Z_t)(\sum_{t=1}^{T}Z_t'Z_t)^{-1}(\sum_{t=1}^{T}Z_t'\tilde{x}_t)]^{-1}(\sum_{t=1}^{T}\tilde{y}_t'Z_t)(\sum_{t=1}^{T}Z_t'Z_t)^{-1}(\sum_{t=1}^{T}Z_t'\varepsilon_t) .$$

Similarly to Proposition 2.1, this estimator is consistent under Assumption 2.2. $\text{plim}[\hat{\beta}_{r2sls}] = \beta + C^{-1}0 = \beta$, where $C \equiv [E(\tilde{y}'Z)E(Z'Z)^{-1}E(Z'\tilde{x})]$. Again, by using the central limit theorem, we have:

$$\sqrt{T}(\hat{\beta}_{r2sls} - \beta) \sim N(0, C^{-1}DC^{-1}) ,$$

where $D \equiv E(\tilde{y}'Z)E(Z'Z)^{-1}E(\varepsilon^2 Z'Z)E(Z'Z)^{-1}E(Z'\tilde{y})$. Then, using the homoskedasticity assumption, we have:

$$Avar[\sqrt{T}(\hat{\beta}_{r2sls} - \beta)] = \sigma_\varepsilon^2[E(\tilde{x}'Z)E(Z'Z)^{-1}E(Z'\tilde{x})]^{-1} = \sigma_\varepsilon^2 A^{-1} .$$

Chapter 3

Hicks-neutral and Non-neutral Productivity[1]

3.1 Introduction

The seminal paper of Olley and Pakes (1996) introduced a structural semi-parametric method, the so-called control function approach, to deal with the endogeneity problem encountered in estimating production functions. This class of estimation techniques has been applied in a large number of recent empirical studies.[2] Following Olley and Pakes (1996), recent developments have exclusively focused on the Cobb-Douglas specification, for example, Levinsohn and Petrin (2003), Ackerberg et al. (2006), Wooldridge (2009), and De Loecker (2011).[3] The Cobb-Douglas specification, however, is an extreme restrictive assumption that ignores key features of the economy, in particular, the non-neutrality of productivity improvements (biased technical change).

In general, the neutrality restriction can be relaxed by considering a class of production functions where technical change is non-separable from the productive factor. In particular the non-linearity of CES specifications allows us to study biased technical change. But at the same time the non-linearity together with unobserved technical change increases the difficulties in estimating parameters. In this chapter, I investigate how a CES production function with biased technical change and non-constant returns to scale can be consistently estimated.

Three approaches were used in the literature for estimating a CES production function. The most common estimation method uses the first order conditions of profit maximization. Based on first order conditions, Berndt (1976) provided estimates of the elasticity of substitution which are close to unity. Antràs (2004) showed that Berndt's

[1] This chapter has been circulated under the title "Estimation of the CES Production Function with Biased Technical Change: A Control Function Approach", Chen (2012).

[2] Empirical studies using the Olley-Pakes control function approach include, for example, Javorcik (2004), Konings and Vandenbussche (2008), and De Loecker (2011).

[3] The interested reader is referred to Ackerberg et al. (2007) and Van Beveren (2012).

results are biased toward the Cobb-Douglas specification, because his estimates suffer from spurious regression bias. The second approach uses the Kmenta (1967) approximation to transform the nonlinear CES function into a linear-in-parameter equation in order to facilitate estimation (for example, Thursby and Lovell, 1978). The third one consists of estimating jointly the first order conditions and production function in a system. The origin of this idea can be traced back to the paper of Nerlove (1967) and a recent application is Klump et al. (2007). Chirinko (2008) provides a survey of the recent literature and shows that the elasticity of substitution lies in the range of 0.4 to 0.6. for U.S. economy. However, all these estimation methods have their limits and are not suitable for the purpose of this study. The first order conditions approach is based on optimizing behavior of producers only, while the Kmenta approximation is based on the production function that captures only technology. Based on different aspects of production analysis, the regression models produce divergent results and contribute to the lack of consensus on the value of the substitution elasticity. Compared to single equation approaches, the system estimation is able to provide more efficient estimates of technology parameters by using both aspects of information (production function and first order conditions), see León-Ledesma et al. (2010). However, this estimation strategy does not follow the recent developments that address the input simultaneity problem.

Firm's input decisions are typically related to productivity. Therefore, the Ordinary Least Squares (OLS) estimation suffers from the simultaneity bias. The traditional estimation methods that control the simultaneity bias, include the Instrumental Variables (IV), the fixed effect (Mundlak, 1961) and the dynamic panel (Arellano and Bond, 1991), however are not able to provide satisfactory results in the case of production function estimation, see Van Beveren (2012). Olley and Pakes (1996) have developed an alternative empirical strategy to overcome endogeneity problems. In this chapter, I combine two strands of literature: the one that focus on estimating the CES production function by using traditional methods (for example, Berndt, 1976, Antràs, 2004, Klump et al., 2007 and León-Ledesma et al., 2010) and the one that deals with endogeneity problems by using the semi-parametric estimation method with a Cobb-Douglas specification (for example, Olley and Pakes, 1996, Levinsohn and Petrin, 2003 and De Loecker, 2011). I contribute to the literature by proposing an extension of the Olley-Pakes method for the CES production function with biased technical change, which allows consistent estimation of the degree of returns to scale, the elasticity of substitution, and the bias in technical change. Both information on technology (characterized by production function) and optimizing behavior of producers (characterized by first order conditions) are used to achieve identification.

This study differs from the existing literature in several ways. First, I generalize beyond the Cobb-Douglas specification to a more flexible CES production function with

Hicks-neutral and factor-augmenting productivity shocks. I propose a semi-parametric estimation method that is able to deal with the endogeneity bias caused by the two unobserved productivity shocks. Second, since I have long time series for many sectors, I estimate the model for different periods and for different sectoral groups in order to understand the technology evolution and the intra-industrial distortion. Using data from U.S. manufacturing industries over the period 1958-2005, it transpires that within the class of CES production functions the unitary elasticity of substitution restriction is rejected. I provide estimates of sectors-level returns to scale and elasticity of substitution, which are 0.95 and 0.63, respectively. The estimation results show that the Cobb-Douglas-based estimator generally overestimates the degree of returns to scale. I also find that the degree of returns to scale is diminishing over time and differs across sectors. By using the estimated elasticity of substitution, I recover the growth rate of relative biased technical change.

The remainder of this chapter is organized as follows: I first present the CES production function with biased technical change and some implications in Section 3.2. In Section 3.3, I discuss the control function approach, the identification conditions and the estimation procedures. Empirical results and robustness checks are given and analyzed in Section 3.4. Section 3.5 concludes.

3.2 A CES production function with biased technical change

Before going into the formal econometric analysis, I frame the problems and give the precise definition of notions discussed above. Firstly, I focus on the CES functional specification, then introduce the technical change terms.

3.2.1 CES specification

Consider a production function $F(.)$ of two factors, labor (L) and capital stock (K) with the value-added output, Y. The elasticity of substitution σ between capital and labor is defined by the percentage change in factor proportions due to a change in the relative marginal products, see Hicks (1932):

$$\sigma \equiv -\frac{\text{dlog}(K/L)}{\text{dlog}(F_K/F_L)} \geq 0, \tag{3.1}$$

where F_K and F_L denote $\partial F/\partial K$ and $\partial F/\partial L$, respectively. Given this definition, Arrow et al. (1961) derived an aggregate production technology with Constant Elasticity of Substitution (CES):

$$Y = F(K, L) = C[\alpha K^{\frac{\sigma-1}{\sigma}} + (1-\alpha)L^{\frac{\sigma-1}{\sigma}}]^{\frac{\rho\sigma}{\sigma-1}}, \tag{3.2}$$

where C is the constant term. Factors are gross complements in production when $\sigma < 1$ and substitutes when $\sigma > 1$. The CES production becomes Cobb-Douglas when $\sigma = 1$. This function is homogeneous of degree ρ in K and L. For any given value of σ, the functional distribution of income is determined by $\alpha \in [0, 1]$. This distribution parameter also depends on the units in which capital and labor are measured and on an arbitrary normalization point. Klump et al. (2012) emphasized the importance of normalizing the CES function, when it comes to identifying the two terms, C and α. Without normalization, the two parameters (C and α) could be any arbitrary point. Here, I focus only on identifying the parameters σ and ρ.

Given the two factors CES production function above, the parameter ρ only represents the degree of returns to scale in capital and labor, i.e.,

$$\rho \equiv \frac{\partial \log Y}{\partial \log L} + \frac{\partial \log Y}{\partial \log K}. \tag{3.3}$$

When capital and labor are increased, if output increases in the same proportion, i.e., $\rho = 1$, then the technology exhibits constant returns to scale. If output increases less than proportionately, i.e., $\rho < 1$, the technology exhibits decreasing returns to scale. If output increases more than proportionately, i.e., $\rho > 1$, the technology exhibits increasing returns to scale. We also need to be aware of the degree of aggregation under study. Basu and Fernald (1997) and Basu (2008) showed that the estimate of returns to scale varies with the aggregation level, in particular it seems to be smaller in disaggregated data. The estimation results presented in Section 3.4 are obtained from U.S. manufacturing data at the six-digit NAICS level.

3.2.2 Factors-augmenting technical change

Technical change can enter the production function in different ways. The most common choice is the Hicks-neutral technology, i.e., $A_h F(K, L)$, as in the case of Cobb-Douglas production function. Hicks-neutrality implies that technical change does not affect the balance between labor and capital demand. Other economic neutrality conditions are Harrod- and Solow-neutrality assumptions. If technical change is Harrod-neutral, the production function becomes $F(K, B_l L)$, where B_l is the labor-augmenting productivity, an increase in productivity is equivalent to having more labor. If technical change is Solow-neutral, i.e., $F(B_k K, L)$, where B_k is the capital-augmenting productivity, an increase in productivity is equivalent to saving capital. In this chapter, I relax these neutrality assumptions by considering the following CES production function:

$$Y = A_h F(B_h K, B_l L) = A_h [\alpha (B_k K)^{\frac{\sigma-1}{\sigma}} + (1-\alpha)(B_l L)^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma\rho}{\sigma-1}}. \tag{3.4}$$

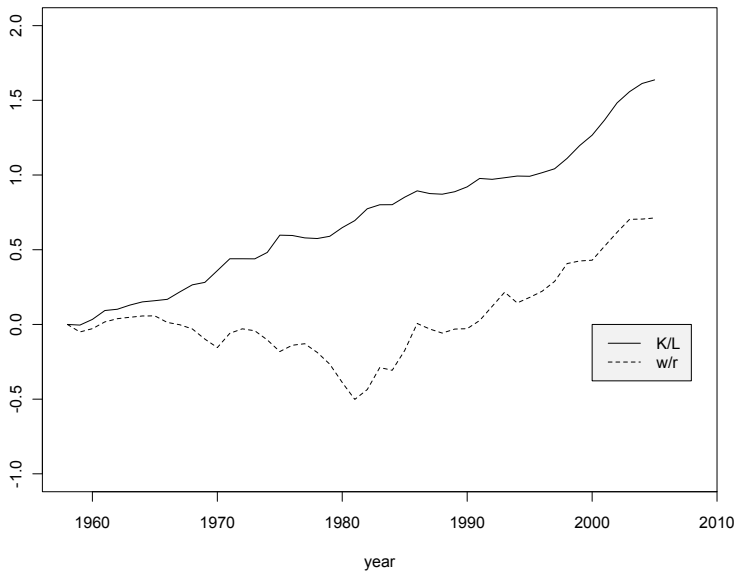


Figure 3.1: Average capital-labor ratio and factor price ratio for U.S. manufacturing industries in the period 1958-2005.

We can also rewrite this production function as:

$$Y = A[\alpha(BK)^{\frac{\sigma-1}{\sigma}} + (1-\alpha)L^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma\rho}{\sigma-1}}, \tag{3.5}$$

where $A \equiv A_h/B_l^{\rho}$ is the relative Hicks-neutral productivity, and $B \equiv B_k/B_l$ is the relative capital-augmenting productivity.

Given a basic assumption that firms minimize costs, firms set marginal products equal to input prices. The first order conditions of the CES production function under cost minimization problem imply that:

$$\frac{K}{L} = \left(\frac{\alpha}{1-\alpha}\right)^{\sigma} \left(\frac{w}{r}\right)^{\sigma} B^{\sigma-1}, \tag{3.6}$$

where w and r denote the wage and the rental capital price, respectively. This equation illustrates that the capital-labor ratio depends on the biased technical change but not on the neutral technical change. If factors are complements in production ($\sigma < 1$), firms reduce their capital-labor ratio when they face an increase in relative capital-augmenting productivity. If factors are substitutes ($\sigma > 1$), firms raise their capital-labor ratio. When $\sigma = 1$ (Cobb-Douglas specification), the effect of biased technical change vanishes and the factors ratio becomes proportional to w/r.

Given any value of the elasticity of substitution, the growth of the capital-labor ratio can be decomposed into two parts, the relative price effects and the biased technical change effects. Figure 3.1 shows the evolution of average capital-labor ratio for U.S.

manufacturing industries in the period 1958-2005.[4] By examining this figure, we can expect that σ was significantly different from one in the period of the late 60s and early 80s. The next section presents the strategy for estimating the parameters of interest, ρ, σ and the growth rate of technical change by using the control function approach.

3.3 Identification and estimation via control functions

The assumption that data reflect technology and optimizing behavior of producers implies that the DGP can be represented by a set of equations, which includes the production function (technology) and the optimal input demand functions (optimization behavior). Both aspects of information are used for identification, in particular the two equations of our regression model are Equations (3.5) and (3.6).

$$\log Y_{it} = \rho \log L_{it} + \frac{\rho\sigma}{\sigma - 1} \log \left[\alpha (B_{it} \frac{K_{it}}{L_{it}})^{\frac{\sigma-1}{\sigma}} + (1 - \alpha) \right] + \log A_{it} + \varepsilon_{it}, \tag{3.7}$$

Equation (3.7) is the logarithmic transformation of (3.5) with an error term appended; $i = 1, ..., N$ indexes sectors and $t = 1, ..., T$ indexes time. The parameters ρ and σ are our central parameters of interest. The scalar disturbance term ε_{it} is an *ex post* shock, which captures the exogenous shocks that are not anticipated by firms. Hence ε_{it} does not affect the optimal choice of labor demand and capital-labor ratio. The endogenous variables L_{it} (optimal labor demand) and $\frac{K_{it}}{L_{it}}$ (optimal capital-labor ratio) are partially determined by unobserved productivity shocks A_{it} and B_{it}. Similar models have been studied by Chesher (2003), Imbens (2007), Imbens and Newey (2009) in the nonparametric framework. Imbens and Newey (2009) provide various partial identification results for the structural equation via the control function approach (e.g. average derivatives, bounds for quantile and average structural function). However, these results are not sufficient for recovering the two technology parameters ρ and σ. The principal reason for the lack of point identification is that the unobserved variable B_{it} is not additively separable from the regressors in the production function. Therefore, in the following lines, I will linearize Equation (3.7) in order to obtain a more tangible form for the empirical investigation.

Firstly, I eliminate the constant term and the potential individual effect by first-differencing model (3.7):

$$\triangle \log Y_{it} = \rho \triangle \log L_{it} + \frac{\rho\sigma}{\sigma - 1} \log \left[\frac{\alpha (B_{it} \frac{K_{it}}{L_{it}})^{\frac{\sigma-1}{\sigma}} + (1 - \alpha)}{\alpha (B_{it-1} \frac{K_{it-1}}{L_{it-1}})^{\frac{\sigma-1}{\sigma}} + (1 - \alpha)} \right] + \triangle \log A_{it} + \triangle \varepsilon_{it}. \tag{3.8}$$

[4] This series is obtained by averaging the 462 U.S. manufacturing industries.

Consider the optimal capital-labor ratio equation:

$$\frac{K_{it}}{L_{it}} = \left(\frac{\alpha}{1-\alpha}\right)^{\sigma} \left(\frac{w_{it}}{r_{it}}\right)^{\sigma} B_{it}^{\sigma-1}, \tag{3.9}$$

where the input price ratio $\frac{w_{it}}{r_{it}}$ is observed and exogenous w.r.t. A_{it} and B_{it} (firms are assumed to be price-takers). We can use (3.9) to substitute the unobservable productivity shock B_{it} from Equation (3.8). Some algebraic manipulation yields:

$$\log(1-\alpha) + \log S_{it} = \log \underbrace{\left[\alpha(B_{it}\frac{K_{it}}{L_{it}})^{\frac{\sigma-1}{\sigma}} + (1-\alpha)\right]}_{S_{it}^*}, \tag{3.10}$$

where the observed variable S_{it} is defined as $\frac{r_{it}K_{it}}{w_{it}L_{it}} + 1$ and S_{it}^* denotes a latent variable. According to (3.10), one can replace the latent variable with the observed one, but as in practice the substitution of the latent variable is usually not perfect, I introduce a scalar measurement error term. This leads to a fully additive regression model:

$$\triangle \log Y_{it} = \rho \triangle \log L_{it} + \gamma \triangle \log S_{it}^* + \triangle a_{it} + \triangle \varepsilon_{it}; \tag{3.11}$$

$$\triangle \log S_{it} = \triangle \log S_{it}^* + \eta_{it}, \tag{3.12}$$

where $a_{it} \equiv \log A_{it}$, $\gamma \equiv \frac{\rho\sigma}{\sigma-1}$ and η_{it} is a classical (zero-mean and uncorrelated with $\triangle \log S_{it}^*$) measurement error term.

The model (3.11)-(3.12) can be viewed as an EIV (Error-in-Variable) model in which we are still facing two endogeneity problems. The first endogeneity problem is due to the Hick-neutral productivity shock that affects the optimal labor demand decision of firms, then the regressor $\triangle \log L_{it}$ is correlated with the unobserved term $\triangle a_{it}$. The second endogeneity problem is that the regressor $\triangle \log S_{it}$ is correlated with the measurement error η_{it}. There are different ways of estimating the parameters in linear regression models with endogenous regressors. 2SLS is one of the most common estimators in linear IV regressions. An asymptotically-equivalent alternative is the control function approach. In this chapter, I use the latter approach because of the lack of valid instruments for controlling productivity shocks. The following sections discuss the control function approach, the identification conditions and the estimation procedures. I also compare the estimations based on different production function specifications (CES versus Cobb-Douglas).

3.3.1 Control function approach

The control function approach was first developed and applied to correct the selection bias of binary response models in Heckman and Robb (1985). This method has been extended for identification of a wide class of models where the explanatory observed

variables and the explanatory unobserved variables are not independently distributed. For instance, it is used for triangular simultaneous equations models in Imbens and Newey (2009), for treatment effect models in Heckman and Vytlacil (2007) and for measurement error models in Hahn et al. (2008). The use of the control function approach for estimating a production function was introduced by Olley and Pakes (1996). The control function approach is a closely related alternative to the classical IV method. In 2SLS estimation, the exogenous variations of instruments are used directly for constructing moment conditions, while the idea of the control function approach is to use control variables (either observed or estimated) that purge the dependence between the observed and unobserved explanatory variables.

Formally, consider a general regression model, $y = f(x, u)$, where the regressor x is correlated with the error term u. Given the assumption that x and u are independent conditionally on a control variable v, Imbens and Newey (2009) give a set of the identification results for this nonlinear models with non-separable disturbances. The identification power of the control variable (v) can be illustrated by the following equation (Imbens and Newey, 2009, p.1488). For any integrable function $\Lambda(y)$,

$$\begin{aligned} \mathrm{E}[\Lambda(y) \mid x, v] &= \int \Lambda(f(x,u)) F_{u|x,v} du \\ &= \int \Lambda(f(x,u)) F_{u|v} du = \mathrm{E}[\Lambda(y) \mid v], \end{aligned}$$

where $F_{u|.}$ is the conditional cumulative distribution function (CDF) of u. The identification comes from the fact that the unobserved variable u of the structural equation can be integrated out by conditioning on v. Since in this chapter, we are dealing only with the linear regression model, a weaker assumption is sufficient: x and u are mean-independent conditionally on v, instead of the stochastic independence. For the rest of this chapter, a valid control variable is defined as follows:

Definition 2.1. *A valid control variable is any observable or estimable variable v such that x and u are mean-independent conditionally on v, i.e.,*

$$\mathrm{E}[u \mid x, v] = \mathrm{E}[u \mid v], \tag{3.13}$$

where x and u are not independently distributed.

Now, as a concrete example consider a simple linear regression model:

$$y = \beta x + u, \tag{3.14}$$

with the reduced form equation of x:

$$x = g(z, v), \tag{3.15}$$

where the variable z is assumed to be a valid instrument that is highly correlated with x and uncorrelated with u and v. The endogeneity of x arises if and only if u is correlated with v. I assume that v is observable or estimable and can be used as a proxy for u:

$$u = m(v) + e, \tag{3.16}$$

where $E[ve] = 0$. Plugging (3.16) into the linear regression model (3.14) gives:

$$y = \beta x + m(v) + e, \tag{3.17}$$

where the function of v is viewed as an additional regressor. It can be shown that v is a valid control variable, which satisfies (3.13), i.e., $\mathrm{E}[u \mid x, v] = \mathrm{E}[u \mid g(z, v), v] = \mathrm{E}[u \mid z, v] = \mathrm{E}[u \mid v]$. The identification of β is achieved in the model (3.17) if and only if x is not an deterministic function of v. Otherwise, there is a collinearity problem, i.e., $y = \beta g(v) + m(v) + e$. Thus, in this setup the identification requires the presence of at least one exogenous variable z in (3.15), which satisfies $\mathrm{E}[u \mid z, v] = \mathrm{E}[u \mid v]$. This is similar to the rank condition in the 2SLS estimation. Since x and v are uncorrelated with the error term e, the parameter β and the function $m(.)$ can be consistently estimated by using the Robinson (1988) estimator. Compared to the 2SLS estimator, the main advantage of using the control function approach in linear regression models is that this estimation method can be implemented whether the instrument z is observed or unobserved.

When instrument z is not available in the data but several candidates of the control variable are observed (typical choices of the control variable for controlling productivity shocks, are the investment, Olley and Pakes, 1996 and the material demand, Levinsohn and Petrin, 2003), then the model (3.17) can be directly estimated. While when z is observed but v is unobserved, the control variable is estimated by using x and z in the first stage of the estimation. For instance, the estimated conditional CDF of x given z has been proved to be a valid control variable, see Imbens and Newey (2009).

3.3.2 Identification conditions

We return to our model of interest, the CES-based model (3.11)-(3.12). The two potential sources of bias are the endogeneity caused by the unobserved productivity shock, a_{it}, and by the measurement error, η_{it}. For controlling the first endogeneity problem (caused by the productivity shock), it seems hard to find a valid instrument but control variables are available. For controlling the second endogeneity problem (caused

by the measurement error), valid instruments are available. Thus, the treatments of the two endogeneity problems are different. Now, I provide conditions that guarantee identification of our model.

Assumption 3.1. *For any observation (indexed by i and t), there is a variable V_{1it} such that the labor demand L_{it} can be written as:*

$$L_{it} = L(Z_{1it}, V_{1it}), \tag{3.18}$$

where a_{it} is mean-independent of Z_{1it} given V_{1it}.

Under Assumption 3.1 the dependence between regressors and the unobserved productivity shock a_{it} can be purged by conditioning on variable V_{1it}. The identification of the model (via control variable V_{1it}) is achieved as long as there is some exogenous variation in either L_{it} or V_{1it}. To see this point, note that:

$$\mathrm{E}[a_{it} \mid L_{it}, V_{1it}] = \mathrm{E}[a_{it} \mid L(Z_{it}, V_{1it}), V_{1it}] = \mathrm{E}[a_{it} \mid Z_{1it}, V_{1it}] = \mathrm{E}[a_{it} \mid V_{1it}], \tag{3.19}$$

where the last equality follows from the mean-independence condition. The choice of control variables depends essentially on whether it satisfies Assumption 3.1. Given an observed control variable V_{1it}, one can replace a_{it} with a nonparametric function of V_{1it}. In a similar way, the second endogeneity problem (caused by the measurement error, η_{it}) can be solved by using the next assumption:

Assumption 3.2. *For any observation, at least one valid instrument is available such that:*

$$\triangle \log S_{it} = h(Z_{2it}) + V_{2it}, \tag{3.20}$$

where η_{it} is mean-independent of Z_{2it} given V_{2it}.

The additivity restriction is imposed in (3.20) for simplifying the estimation procedure. Given the data at hand, V_{2it} is not observed. However, we can assume that the growth rate of $\frac{w_{it}}{r_{it}}$ or the lagged values of S_{it} are valid instruments (Z_{2it}) and estimate V_{2it} in the first-stage. The traditional approach in the linear EIV case is to estimate the model by using the 2SLS estimator. But the estimation procedure based on the control function approach is more convenient here. Under Assumption 3.2, the residuals V_{2it} is a valid control variable:

$$\mathrm{E}[\eta_{it} \mid \triangle \log S_{it}, V_{2it}] = \mathrm{E}[\eta_{it} \mid h(Z_{2it}) + V_{2it}, V_{2it}] = \mathrm{E}[\eta_{it} \mid Z_{2it}, V_{2it}] = \mathrm{E}[\eta_{it} \mid V_{2it}], \tag{3.21}$$

and by construction V_{2it} is a proxy for η_{it}, then the model (3.11)-(3.12) can be identified

by inverting out η_{it}.

3.3.3 Estimation procedures

Now I describe the estimation procedure for the CES based model (3.11)-(3.12), which follows closely the previous identification discussion. I also review briefly the estimation strategy proposed by Olley and Pakes (1996) for the Cobb-Douglas model, which is included in our empirical studies for comparison.

The Cobb-Douglas based model (Olley and Pakes, 1996) The regression model based on the Cobb-Douglas production function is:

$$\triangle \log Y_{it} = \beta_l \triangle \log L_{it} + \beta_k \triangle \log K_{it} + \triangle a_{it} + \triangle \varepsilon_{it}. \tag{3.22}$$

Olley and Pakes (1996) assume that the investment function $I_t(.)$ is strictly monotonic in a_{it}, i.e.,

$$I_{it} = I_t(K_{it}, a_{it}). \tag{3.23}$$

Then (3.22) can be inverted to $a_{it} = I_t^{-1}(K_{it}, I_{it})$, where the variables K_{it} and I_{it} are used as control variables. Substituting this inverse function into the model (3.22), we have:

$$\triangle \log Y_{it} = \beta_l \triangle \log L_{it} + \Phi(K_{it}, I_{it}, K_{it-1}, I_{it-1}) + \triangle \varepsilon_{it}, \tag{3.24}$$

where $\Phi(K_{it}, I_{it}, K_{it-1}, I_{it-1}) = \beta_k \triangle \log K_{it} + I_t^{-1}(K_{it}, I_{it}) - I_{t-1}^{-1}(K_{it-1}, I_{it-1})$. Thus, we can estimate the parameter β_l and the nonparametric functions by applying Robinson's (1988) estimator on (3.24). Olley and Pakes (1996) assume that productivity a_{it} evolves exogenously as a first-order Markov process:

$$\begin{aligned} a_{it} &= E[a_{it} \mid information_{t-1}] + \xi_{it} \\ &= E[a_{it} \mid a_{it-1}] + \xi_{it}, \end{aligned}$$

where a firm's expectations about future productivity depend only on a_{it-1}, and ξ_{it} is the unexpected innovation in a_{it} that is orthogonal to the information set at $t-1$, i.e., $E[\xi_{it} \mid information_{t-1}] = 0$. Assuming a linear model $a_{it} = \tau + \rho a_{it-1} + \xi_{it}$, the implied innovation term is $\xi_{it} = a_{it} - \tau - \rho a_{it-1}$. Given the estimated coefficient $\hat{\beta}_l$ of the first-stage estimation, we can rewrite the unexpected innovation as a parametric function of observations:

$$\xi_{it}(\beta_k, \tau, \rho) = \log Y_{it} - \hat{\beta}_l \log L_{it} - \beta_k \log K_{it} - \tau - \rho(\log Y_{i-1} - \hat{\beta}_l \log L_{it-1} - \beta_k \log K_{it-1}).$$

Typically, the production timing assumption suggests that the period s capital demand (K_{is}) and the period $s-1$ labor demand (L_{is-1}), for $s \leq t$, are decided upon

at $t-1$ or before, and are included in the information set at $t-1$. This assumption implies that both K_{is} and L_{is-1} must be uncorrelated with the unexpected innovation (ξ_{it}). In particular, I estimate the parameter β_k, τ and ρ by using GMM with the following instruments: the capital accumulation at t ($\triangle \log K_{it}$) and the lagged value of labor demand ($\log L_{it-1}$).[5] Given the estimates of β_l and β_k, the degree of returns to scale is computed as the sum of the two estimated parameters, $\hat{\rho}_{CD} = \hat{\beta}_l + \hat{\beta}_k$ under the Cobb-Douglas specification. The estimator $\hat{\rho}_{CD}$ is a sequential two-step estimator whose standard error is computed using the bootstrap.

The CES based model The treatment of the unobserved productivity a_{it} remains the same as in the Olley-Pakes estimation method, but the additional difficulty here is that the regressor $\triangle \log S_{it}$ is correlated with the measurement error η_{it}. Consider the model (3.11)-(3.12) where the term $\triangle a_{it}$ is substituted as before:

$$\triangle \log Y_{it} = \rho \triangle \log L_{it} + \gamma \triangle \log S_{it} - \gamma \eta_{it} + \varphi(K_{it}, I_{it}, K_{it-1}, I_{it-1}) + \triangle \varepsilon_{it}, \tag{3.25}$$

with $\varphi(K_{it}, I_{it}, K_{it-1}, I_{it-1}) = I_t^{-1}(K_{it}, I_{it}) - I_{t-1}^{-1}(K_{it-1}, I_{it-1})$. If the control variable is directly observed, the measurement error term η_{it} can be proxied by a nonparametric function, i.e., $E[\gamma \eta_{it} \mid V_{2it}] = \Gamma(V_{2it})$. However, the control variable V_{2it} is unobserved, then it has to be estimated first. Assume that the growth rate of input price ratio, i.e., $Z_{2it} = \frac{w_{it}}{r_{it}}$ is a valid instrument that satisfies Assumption 3.2. The preliminary estimated values of the control variable V_{2it} are obtained based on the kernel estimation: $\hat{V}_{2it} = \triangle \log S_{it} - \hat{h}(Z_{2it})$. Since we cannot insert an estimated variable into a nonlinear function, we need to restrict the nonparametric function $\Gamma(.)$ to a linear parametric one, i.e., $E[\gamma \eta_{it} \mid V_{2it}] = -\theta V_{2it}$.

$$\triangle \log Y_{it} = \rho \triangle \log L_{it} + \gamma \triangle \log S_{it} + \theta \hat{V}_{2it} + \varphi(K_{it}, I_{it}, K_{it-1}, I_{it-1}) + \triangle \varepsilon_{it}. \tag{3.26}$$

Therefore, the parameter ρ, γ and θ can be consistently estimated by using the Robinson (1988) estimator. Since the first-stage estimation is nonparametric, the asymptotic variance matrix of the estimator depends on preliminary estimates and it is difficult to compute using general results for semi-parametric regression models. Thus, the corresponding standard errors are obtained using the bootstrap. Henceforth, the estimated elasticity of substitution is recovered as: $\hat{\sigma} = \frac{\hat{\gamma}}{\hat{\gamma} - \hat{\rho}}$.

3.3.4 Bias of the Cobb-Douglas specification

Now the question is, what differences should we expect in term of estimation outcomes between the Cobb-Douglas-based regression model and the CES-based regression model.

[5] The starting values of GMM estimation is: $c = 0$, $\rho = 1$ and $\hat{\beta}_k^{ols}$ obtained by regressing $\triangle \hat{a}_{it} \equiv \triangle \log Y_{it} - \hat{\beta}_l \triangle \log L_{it} - \beta_k \triangle \log K_{it}$ on $\triangle \log K_{it}$.

Firstly, the Cobb-Douglas model sets the elasticity of substitution, σ, to one. Thus, if the economy was not characterized by the Cobb-Douglas technology, then we should find estimates of σ that significantly differ from one. Secondly, both models produce estimates of returns to scale, ρ. I will show in the following lines that the Cobb-Douglas specification may overestimates the returns to scale.

For the sake of clarity and convenience, I focus only on the bias of misspecification. Thus, the technical change terms, A_{it} and B_{it} are disregarded in this subsection. Consider the Kmenta approximation of (3.2):

$$\log Y_{it} = C + \rho\alpha \log K_{it} + \rho(1-\alpha)\log L_{it} + \frac{1}{2}\rho\frac{\sigma-1}{\sigma}\alpha(1-\alpha)(\log K_{it} - \log L_{it})^2, \quad (3.27)$$

where C is the constant term. The last term of the right hand side is ignored when one considers the Cobb-Douglas production function. The first-difference transformation yields:

$$\begin{aligned} \triangle \log Y_{it} = \rho\alpha \triangle \log K_{it} + \rho(1-\alpha)\triangle \log L_{it} \quad (3.28) \\ + \tfrac{1}{2}\rho\tfrac{\sigma-1}{\sigma}\alpha(1-\alpha)[(\log K_{it} - \log L_{it})^2 - (\log K_{it-1} - \log L_{it-1})^2]. \end{aligned}$$

If $\hat{\beta}_k$ and $\hat{\beta}_l$ are the estimated coefficients of $\triangle \log K$ and $\triangle \log L$ based on the Cobb-Douglas specification. According to the well known results of Theil (1957), the expectations of these estimators are:

$$E(\hat{\beta}_k) = \rho\alpha + \frac{1}{2}\rho\frac{\sigma-1}{\sigma}\alpha(1-\alpha)\hat{\pi}_k; \quad (3.29)$$

$$E(\hat{\beta}_l) = \rho(1-\alpha) + \frac{1}{2}\rho\frac{\sigma-1}{\sigma}\alpha(1-\alpha)\hat{\pi}_l, \quad (3.30)$$

where $\hat{\pi}_k$ and $\hat{\pi}_l$ are the two estimates obtained from the regression of the omitted variable, $(\log K_{it} - \log L_{it})^2 - (\log K_{it-1} - \log L_{it-1})^2$, on the included variables, $\triangle \log K_{it}$ and $\triangle \log L_{it}$. Thus, the bias of the estimated returns to scale by using the Cobb-Douglas specification is:

$$E(\hat{\beta}_k + \hat{\beta}_l) - \rho = \frac{1}{2}\rho\frac{\sigma-1}{\sigma}\alpha(1-\alpha)(\hat{\pi}_k + \hat{\pi}_l). \quad (3.31)$$

Given the parameter ρ is positive and $\alpha \in [0,1]$, the bias of estimated returns to scale based on the Cobb-Douglas specification are summarized in the following table.

The case of negative elasticity $\sigma < 0$ is not allowed by economic theory, but due to estimation errors this case is empirically possible. Since the empirical results in this chapter suggest that $\sigma < 1$ (see Section 3.4) and $\hat{\pi}_k + \hat{\pi}_l$ is generally negative, we can expect that the regression based on the Cobb-Douglas specification overestimates the degree of returns to scale. The estimation results in the next section will confirm this conclusion.

Table 3.1: Bias of estimated returns to scale based on the Cobb-Douglas specification

	$\hat{\pi}_k + \hat{\pi}_l < 0$	$\hat{\pi}_k + \hat{\pi}_l = 0$	$\hat{\pi}_k + \hat{\pi}_l > 0$
$\sigma < 1$	overestimation	unbiased	underestimation
$\sigma = 1$	unbiased	unbiased	unbiased
$\sigma > 1$	underestimation	unbiased	overestimation
$\sigma < 0$	underestimation	unbiased	overestimation

Table 3.2: Estimates of the full panel (1958-2005)

	Cobb-Douglas	CES
Labor (β_l)	0.907 (0.008)	-
Capital (β_k)	0.306 (0.016)	-
Returns to Scale (ρ)	1.213 (0.019)	0.954 (0.025)
Elasticity of Substitution (σ)	1	0.629 (0.009)
$\pi_k + \pi_l$	-0.524 (0.033)	

3.4 Empirical investigation

The empirical investigation focuses on U.S. manufacturing industries at six-digit NAICS aggregation level. The information needed for conducting the econometric analysis comes from the NBER Manufacturing Industry database, which contains annual information on output, employment, payroll, investment, capital stock and other inputs cost together with prices deflators of 462 industries from 1958 to 2005. The construction of this database has been discussed in the technical report of Bartlesman and Gray (1996). The detailed description of this data set is reported in Appendix. Compared to firm-level data sets, the NBER data set offers some advantages. Firstly, it contains the price indexes that are the essential information for characterizing the optimizing behavior. Secondly, it allows us to avoid the multiple products problem of the firm-level data. Finally, at the six-digit NAICS aggregation level we still have a large number of sectors, which guarantees a good asymptotic approximation for cross-sectional regressions.

3.4.1 Estimation results

I start by reporting the estimates of returns to scale and elasticity of substitution for different windows of observation and for different sector groups. Then, given the estimates of technology parameters, I recover the Hicks-neutral and the factor-augmenting productivity and compute their annual growth rates.

Table 3.2 summarizes the estimation results over the full panel as well as the estimated standard errors (obtained by using the bootstrap with 1000 replications). The second column reports the estimates of parameters β_l and β_k based on the Cobb-Douglas specification (following Olley and Pakes, 1996). The third column gives the estimation

results for the CES model. The degree of returns to scale defined in (3.3) is computed as the sum of β_l and β_k in the Cobb-Douglas model, which is 1.213 with a 95% confidence interval [1.176, 1.250]. This result indicates that the industries were characterized by increasing returns to scale technology. The estimated degree of returns to scale obtained from the CES based model is 0.954 with a 95% confidence interval [0.906, 1.003], which suggests that the technology exhibits non-increasing returns to scale. The estimated elasticity of substitution is 0.629 with a 95% confidence interval [0.611, 0.647] that is far from covering one. In Section 3.3.4, I showed that the Cobb-Douglas based estimation of returns to scale suffers from an omitted variables bias when the elasticity of substitution differs from unity, see Table 3.1. The estimate of $\pi_k + \pi_l$ in Equation (3.31) is negative and significantly different from zero, which indicates that the Cobb-Douglas based regression overestimates the degree of returns to scale.

When T is large in the panel, one potential concern is the non-stationarity of the data. The first-difference transformation could stationarize series in the linear function, but not for nonlinear parts of the model. Therefore, given the non-stationarity the question we need to ask is whether the estimation results obtained by using the long panels are misleading? To answer this, I consider shorter panels, where $T = 3$ and compare the estimation results with previous findings. In this case, the estimation relies mainly upon the cross-sectional variation, thus the results are less affected by the problem of non-stationarity. The estimation results are reported in Table 3.3. The evolution of estimated returns to scale and elasticity of substitution with the 95% confidence intervals are depicted in Figures 3.2 and 3.3, respectively.

On the average, I find the similar estimation results as for the full panel case. The average estimates obtained from the CES based model suggest that the industries were characterized by a non-increasing returns to scale technology with the non-unitary elasticity of substitution, while the Cobb-Douglas based model predicts increasing returns to scale. The average estimated returns to scale obtained from the Cobb-Douglas specification and the CES specification are 1.164 and 0.819, respectively. The average estimated elasticity of substitution is 0.675.

Comparing the estimates of returns to scale obtained from the two models, we see that the Cobb-Douglas based regressions overestimate the degree of returns to scale in the majority of cases (14 out of 16 panels). Figures 3.2 and 3.3 show that the estimates of returns to scale are diminishing over time in both models, while the estimates of the elasticity of substitution are relatively stable. By regressing the estimates of returns to scale on a linear trend, I find that the decreasing rates are 3.4% (based on the CES specification) and 3.9% (based on the Cobb-Douglas specification) for each period of three years. This result may reflect the fact that the growth of U.S manufacturing industries was more and more driven by the technical change rather than the economies of scale.

From Figure 3.3, we can see that the confidence intervals of the estimated elasticity of substitution lay entirely below one for 8 out of 16 panels. In 3 out of 16 panels, the confidence intervals cover one, where we cannot conclude on the substitutability of production factors. The estimated elasticity of substitution should not be heeded in 5 cases, because their standard errors are very large. For the cases in which the estimated elasticity of substitution are significantly below unity, the estimates of $\pi_l + \pi_k$ predict correctly the bias of estimated returns to scale based on the Cobb-Douglas specification. For example, in the panel "64-65-66", the estimated $\pi_l + \pi_k$ is significantly negative and the Cobb-Douglas based regression overestimates the degree of returns to scale; in the panel "88-89-90", the estimated $\pi_l + \pi_k$ is not significantly different from zero and the two estimates of returns to scale are close.

The size of elasticity of substitution has important economic implications, for example σ is critical for determining the pattern of capital accumulation or the path of growth. Previous estimation procedures only produce the economy's aggregated estimates. The elasticity of substitution, however, may differ across sectors. Now I stratify the panel according to the sectoral classification (the 3-digit NAICS) and perform the regressions for each sub-group of manufacturing industry. The estimation results are reported in Table 3.4. Figures 3.4 and 3.5 depict the corresponding estimates with 95% confidence intervals for different sectors.

There are significant differences among the estimates of technology parameters across the sectors. The estimates of returns to scale lay in the range of 0.566 to 1.173 with the CES model, and the estimates of the elasticity of substitution lay in the range of 0.479 to 0.865. As for previous findings, the Cobb-Douglas based regressions overestimate the degree of returns to scale in the majority of sectors, expect for Sector 316 (Leather & allied prod) where the estimated returns to scale by considering the CES specification is higher. All estimates of the elasticity of substitution are significantly below one, which rejects once again the Cobb-Douglas specification. In 15 out of 20 cases, the estimates of $\pi_l + \pi_k$ are negative and significantly different from zero, which explain the overestimation of returns to scale by the Cobb-Douglas based regression. In other cases, the estimates of $\pi_l + \pi_k$ have relatively large estimated standard errors that we cannot conclude on the direction of the bias for the Cobb-Douglas based regression.

Table 3.3: Estimates with the short panel of 3 periods

	Cobb-Douglas			CES		
Panels	β_l	β_k	$\beta_l + \beta_k$	ρ	σ	$\pi_k + \pi_l$
58-59-60	1.066 (0.032)	0.258 (0.076)	1.324 (0.087)	1.237 (0.064)	0.377 (3.790)	−0.959 (0.108)
61-62-63	1.019 (0.027)	0.623 (0.088)	1.642 (0.090)	1.060 (0.074)	0.847 (0.045)	−0.051 (0.137)
64-65-66	0.958 (0.033)	0.478 (0.151)	1.436 (0.144)	1.001 (0.073)	0.819 (0.041)	−0.245 (0.092)
67-68-69	0.981 (0.043)	0.383 (0.098)	1.364 (0.101)	1.107 (0.078)	0.861 (0.023)	−0.519 (0.125)
70-71-72	0.783 (0.029)	0.816 (0.112)	1.600 (0.115)	0.860 (0.127)	0.910 (0.075)	−0.236 (0.193)
73-74-75	0.972 (0.044)	−0.486 (0.148)	0.487 (0.156)	0.766 (0.076)	0.837 (0.035)	0.125 (0.172)
76-77-78	0.903 (0.037)	0.408 (0.128)	1.310 (0.126)	0.831 (0.094)	0.622 (0.193)	0.386 (0.147)
79-80-81	1.137 (0.051)	0.091 (0.096)	1.228 (0.112)	0.965 (0.103)	0.685 (0.078)	−0.652 (0.115)
82-83-84	0.971 (0.048)	0.394 (0.127)	1.365 (0.127)	0.551 (0.122)	0.691 (0.072)	−0.300 (0.134)
85-86-87	0.848 (0.040)	0.340 (0.132)	1.188 (0.133)	0.716 (0.126)	−0.266 (11.247)	−0.524 (0.183)
88-89-90	0.762 (0.049)	0.095 (0.117)	0.856 (0.122)	0.735 (0.136)	0.767 (0.067)	0.198 (0.142)
91-92-93	0.716 (0.037)	0.325 (0.148)	1.041 (0.151)	0.478 (0.089)	0.829 (0.449)	0.626 (0.124)
94-95-96	0.605 (0.053)	0.376 (0.197)	0.981 (0.207)	0.601 (0.186)	0.905 (0.234)	0.256 (0.133)
97-98-99	0.833 (0.053)	0.225 (0.100)	1.057 (0.142)	0.622 (0.148)	0.656 (0.092)	−0.461 (0.170)
00-01-02	0.803 (0.041)	0.121 (0.152)	0.924 (0.175)	1.045 (0.169)	0.812 (5.444)	0.169 (0.227)
03-04-05	0.580 (0.054)	0.251 (0.134)	0.831 (0.123)	0.532 (0.198)	0.451 (37.401)	−0.450 (0.131)
Mean	0.871	0.293	1.164	0.819	0.675	-

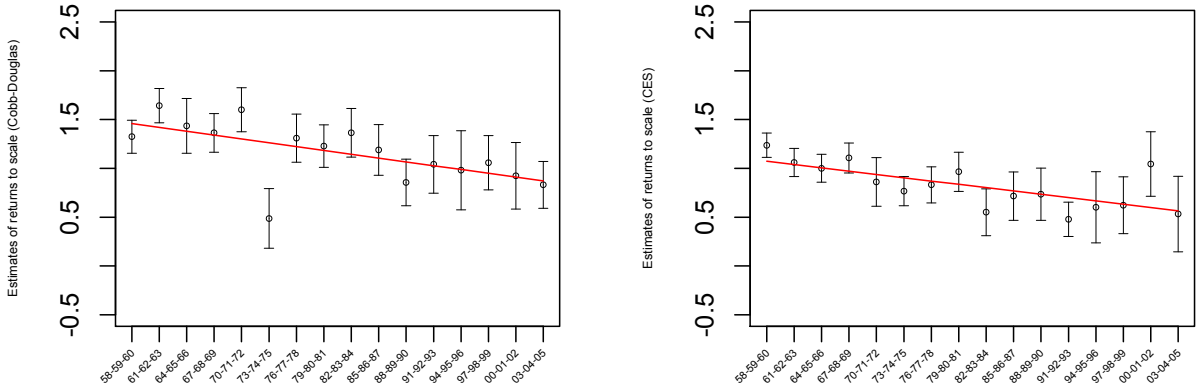


Figure 3.2: Estimates of returns to scale with 95% confidence intervals
Note: the sloping line represents the fitted line of the regression of estimates on a linear trend.

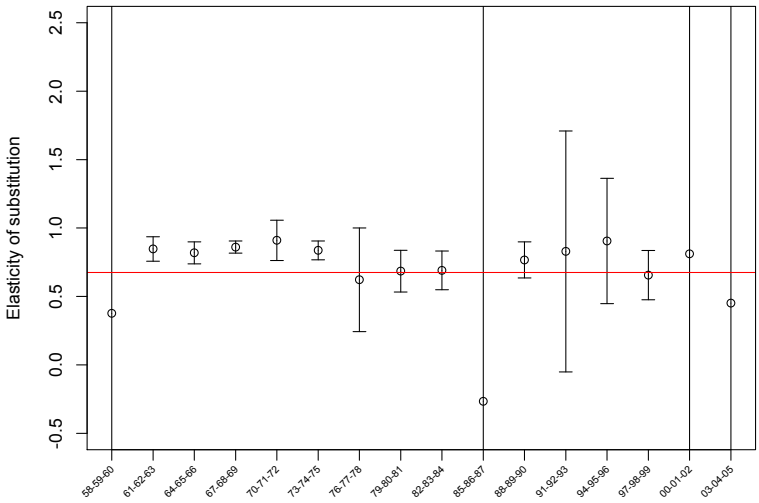


Figure 3.3: Estimates of elasticity of substitution with 95% confidence intervals
Note: the horizontal line represents the average value of estimates when the panels "59-59-60", "64-65-66", "67-68-69", "00-01-02" and "03-04-05" are disregarded.

Table 3.4: Estimates with sectoral stratification

		Cobb-Douglas			CES		
Sectors	$N \cdot T$	β_l	β_k	$\beta_l + \beta_k$	ρ	σ	$\pi_k + \pi_l$
Food mfg (311)	2112	0.809 (0.042)	0.384 (0.112)	1.242 (0.127)	0.767 (0.121)	0.546 (0.048)	−2.108 (0.100)
Beverage & tobacco prod. (312)	432	0.768 (0.077)	0.191 (0.136)	1.035 (0.197)	0.785 (0.212)	0.557 (0.106)	−0.789 (0.199)
Textile & mills (313&314)	864	0.820 (0.041)	0.241 (0.085)	1.129 (0.106)	0.566 (0.111)	0.709 (0.055)	−0.592 (0.101)
Apparel (315)	1104	0.838 (0.031)	0.152 (0.066)	1.032 (0.088)	0.940 (0.111)	0.795 (0.082)	−2.847 0.163)
Leather & allied prod.(316)	480	0.847 (0.057)	−0.147 (0.194)	0.713 (0.199)	0.715 (0.164)	0.865 (0.050)	−1.218 (0.221)
Wood product mfg (321)	672	0.794 (0.043)	0.197 (0.120)	1.040 (0.123)	0.791 (0.087)	0.687 (0.042)	0.146 (0.092)
Paper mfg (322)	960	0.742 (0.039)	0.240 (0.126)	1.051 (0.178)	0.579 (0.253)	0.642 (0.058)	−0.638 (0.164)
Printing (323)	576	0.870 (0.024)	0.201 (0.108)	1.122 (0.127)	1.032 (0.110)	0.664 (0.034)	−0.862 (0.106)
Petroleum & coal prod. (324)	240	0.890 (0.125)	0.294 (0.212)	1.230 (0.279)	1.089 (0.358)	0.677 (0.118)	0.160 (0.305)
Chemical mfg (325)	1632	0.814 (0.041)	0.191 (0.145)	1.049 (0.151)	0.598 (0.097)	0.479 (0.074)	−0.422 (0.096)
Plastics & rubber prod. (326)	768	0.967 (0.040)	0.377 (0.097)	1.143 (0.095)	0.806 (0.094)	0.562 (0.048)	−0.341 (0.073)
Nonmetallic mineral prod.(327)	1152	0.952 (0.029)	0.002 (0.071)	1.019 (0.077)	0.994 (0.065)	0.568 (0.029)	0.099 (0.084)
Primary metal mfg (331)	1248	0.966 (0.052)	0.032 (0.145)	1.044 (0.154)	0.961 (0.131)	0.627 (0.051)	−0.669 (0.145)
Fabricated metal prod. (332)	2064	0.925 (0.020)	0.268 (0.114)	1.243 (0.117)	1.029 (0.064)	0.639 (0.024)	0.005 (0.077)
Machinery (333)	2352	1.012 (0.025)	0.153 (0.136)	1.215 (0.138)	1.041 (0.075)	0.613 (0.049)	−0.548 (0.070)
Computer & electro. prod. (334)	1344	0.867 (0.031)	0.679 (0.097)	1.592 (0.101)	1.173 (0.085)	0.479 (0.209)	−1.062 (0.147)
Electrical equipment (335)	1056	0.906 (0.040)	0.204 (0.092)	1.157 (0.094)	0.694 (0.111)	0.680 (0.042)	−0.169 (0.106)
Transportation equipment (336)	1440	1.160 (0.029)	0.031 (0.071)	1.235 (0.075)	0.991 (0.105)	0.669 (0.035)	−0.438 (0.103)
Furniture & related prod.(337)	576	0.856 (0.042)	0.185 (0.125)	1.088 (0.148)	0.788 (0.167)	0.812 (0.054)	−0.344 (0.105)
Miscellaneous (339)	1104	0.770 (0.031)	0.294 (0.101)	1.111 (0.112)	0.812 (0.160)	0.641 (0.083)	−0.740 (0.086)

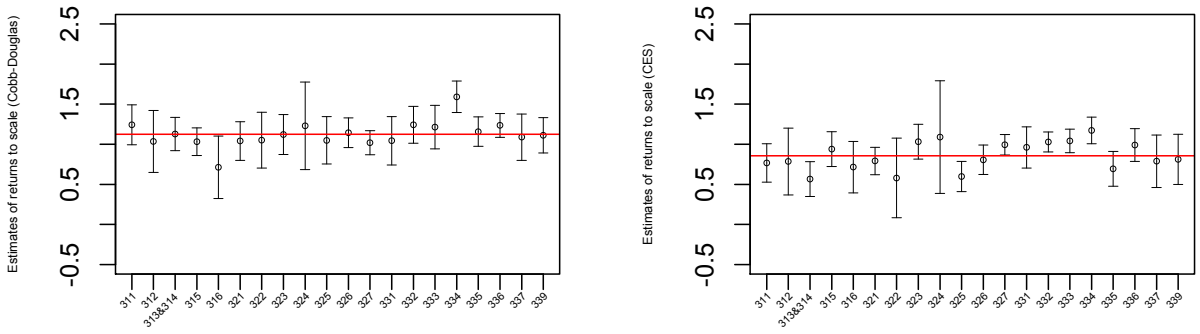


Figure 3.4: Estimates of returns to scale with 95% confidence intervals for different sectors

Note: the horizontal line represents the average value of estimates.

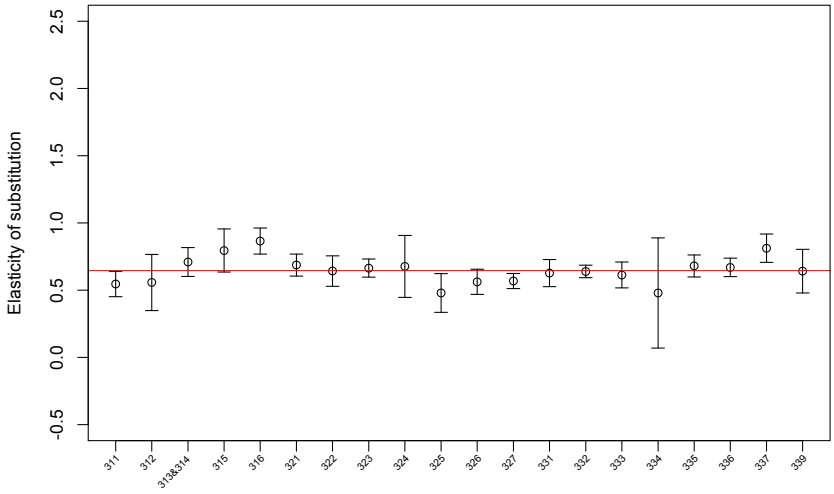


Figure 3.5: Estimates of elasticity of substitution with 95% confidence intervals for different sectors

Note: the horizontal line represents the average value of estimates.

3.4.2 Recovering the Hicks-neutral and factor-augmenting productivity

Given the estimates of technology parameters, I can now recover the relative Hicks-neutral and factor-augmenting productivity in logarithmic forms, $\log A_{it}$ and $\log B_{it}$, and then compute the corresponding growth rates. The term $\log A_{it}$ is defined as $\log A_{it} = \log A_{hit} - \rho \log B_{lit}$ and the term $\log B_{it}$ is defined as $\log B_{it} = \log B_{kit} - \log B_{lit}$, where A_h is the net Hicks-neutral productivity, B_l is the net labor-augmenting productivity and B_k is the net capital-augmenting productivity, see Equation (3.5). Since the net productivity terms cannot be identified separately, we only interpret the productivity measures in relative terms.

Given the estimates obtained from the CES based regression, the relative Hicks-neutral productivity $\log A_{it}$ is recovered by using Equation (3.7) as:

$$\log \hat{A}_{it} + c = \log Y_{it} - \hat{\rho} \log L_{it} - \hat{\gamma} \log S_{it}, \tag{3.32}$$

where $c = \gamma \log(1 - \alpha)$. For the sake of comparison, we can also compute the Hicks-neutral productivity under the Cobb-Douglas model, which is denoted with superscript *CD*.

$$\log \hat{A}_{it}^{CD} = \log Y_{it} - \hat{\beta}_l \log L_{it} - \hat{\beta}_k \log K_{it}. \tag{3.33}$$

Given the estimated elasticity of substitution, I can invert the capital-labor ratio equation (3.9) to obtain the expression of logarithmic relative factor-augmenting productivity:

$$\log \hat{B}_{it} + d = \frac{\hat{\sigma}}{\hat{\sigma} - 1} \log \frac{r_{it}}{w_{it}} + \frac{1}{\hat{\sigma} - 1} \log \frac{K_{it}}{L_{it}}, \tag{3.34}$$

where $d = \frac{\sigma}{\sigma - 1} \log \left(\frac{\alpha}{1 - \alpha} \right)$. As mentioned above the parameter α is not identified under our estimation procedure, but this is not a problem here, these constant terms do not affect the estimation of the growth rate.

Consider the estimates obtained from the full panel cases,[6] after averaging over sectors, I examine the time variation of the aggregated productivity growth rates:

$$\lambda_t^A \equiv N^{-1} \sum_i^N \triangle \log \hat{A}_{it} \quad \text{and } \lambda_t^B \equiv N^{-1} \sum_i^N \triangle \log \hat{B}_{it}.$$

Figure 3.6 compares the estimation of the relative Hicks-neutral productivity growth, λ_t^A and of the relative labor-augmenting productivity growth, $-\lambda_t^B$ obtained from the CES based model for the period 1959-2005.

[6] Under the CES specification, the estimates are: $\hat{\sigma} = 0.629$, $\hat{\rho} = 0.954$, $\hat{\gamma} = 1.617$ and $\hat{\theta} = -1.655$; under the Cobb-Douglas specification, the estimates are: $\hat{\beta}_l = 0.907$ and $\hat{\beta}_k = 0.306$

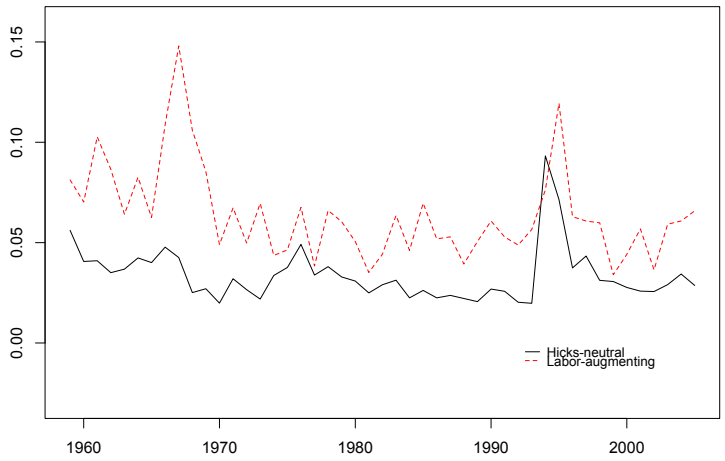


Figure 3.6: Estimation of the relative Hicks-neutral and the relative labor-augmenting productivity growth rates (λ_t^A and $-\lambda_t^B$) for the period 1959-2005

Figure 3.6 shows that both time series are stationary and that the relative labor-augmenting productivity growth rates ($-\lambda_t^B$) is more volatile than the relative Hicks-neutral productivity growth rates (λ_t^A). The time series of the relative labor-augmenting productivity growth consists of two main spikes. The earlier spike was in 1967, and the second one has been in 1995 where a spike of λ_t^A appeared at the same period. The two series are positively correlated with a correlation coefficient of 0.486. Now, by averaging over sectors and periods, I compute the average (annual) productivity growth rates as:

$$\bar{\lambda}^A \equiv T^{-1}N^{-1}\sum_t^T\sum_i^N \triangle \log \hat{A}_{it} \quad \text{and } \bar{\lambda}^B \equiv T^{-1}N^{-1}\sum_t^T\sum_i^N \triangle \log \hat{B}_{it}.$$

I obtain an average Hicks-neutral productivity growth ($\bar{\lambda}^A$) of 3.37% based on the CES model (3.32), while under the Cobb-Douglas model (3.33) the average Hicks-neutral productivity growth is 1.89%. I find that labor-augmenting technical progress grew annually about 6.42% faster than capital-augmenting technical progress, i.e., $\bar{\lambda}^B = \bar{\lambda}^{Bk} - \bar{\lambda}^{Bl} = -6.42\%$, where $\bar{\lambda}^{Bk}$ denotes the net capital-augmenting productivity growth rate and $\bar{\lambda}^{Bl}$ denotes the net labor-augmenting productivity growth rate.[7] Our estimation of $\bar{\lambda}^B$ is larger than the one obtained by Antràs (2004), i.e., 3.15%. But both findings lead to the same conclusion that when the production technology is characterized by $\sigma < 1$, all firms have the incentive to pursue labor-augmenting innovations on the balanced growth path rather than capital-augmenting innovations (the theoretical justifications can be found in Acemoglu, 2003).

[7]Under the Cobb-Douglas specification, by construction, the net factor-augmenting productivity growth is restricted to zero, i.e., $\bar{\lambda}^{Bl} = \bar{\lambda}^{Bk} = 0$.

Table 3.5: Estimates of average productivity growth rates with sectoral heterogeneity

Sectors	$N \cdot T$	ρ	σ	$\bar{\lambda}^A$	$\bar{\lambda}^B$
Food mfg (311)	2112	0.767	0.546	0.037	−0.071
Beverage & tobacco prod. (312)	432	0.785	0.557	0.039	−0.064
Textile & mills (313&314)	864	0.566	0.709	0.035	−0.083
Apparel (315)	1104	0.940	0.795	0.051	−0.223
Leather & allied prod.(316)	480	0.715	0.865	0.023	−0.242
Wood product mfg (321)	672	0.791	0.687	0.025	−0.064
Paper mfg (322)	960	0.579	0.642	0.026	−0.060
Printing (323)	576	1.032	0.664	0.023	−0.072
Petroleum & coal prod. (324)	240	1.089	0.677	0.045	−0.062
Chemical mfg (325)	1632	0.598	0.479	0.035	−0.042
Plastics & rubber prod. (326)	768	0.806	0.562	0.036	−0.056
Nonmetallic mineral prod.(327)	1152	0.994	0.568	0.028	−0.040
Primary metal mfg (331)	1248	0.961	0.627	0.027	−0.050
Fabricated metal prod. (332)	2064	1.029	0.639	0.024	−0.053
Machinery (333)	2352	1.041	0.613	0.023	−0.054
Computer & electro. prod. (334)	1344	1.173	0.479	0.070	−0.072
Electrical equipment (335)	1056	0.694	0.680	0.031	−0.063
Transportation equipment (336)	1440	0.991	0.669	0.028	−0.041
Furniture & related prod.(337)	576	0.788	0.812	0.030	−0.107
Miscellaneous (339)	1104	0.812	0.641	0.031	−0.066

The previous estimation of productivity growth are obtained by assuming that all sectors have the same technology, which is characterized by the degree of returns to scale (ρ) and the elasticity of substitution (σ) under the CES specification. However, the results in Table 3.4 suggest that the production technology may differ across sectors. Table 3.5 summarizes the estimation of (relative) average productivity growth rates by taking into consideration the sectoral heterogeneity. Figure 3.7 displays the estimated values of average productivity growth rates for the 20 sectoral groups. In most cases, the estimates of $\bar{\lambda}^A$ lay in the range of 0.02 to 0.05; the estimates of $\bar{\lambda}^B$ are negative (labor-augmenting technical progress grew faster than capital-augmenting technical progress) and lay in the range of -0.04 to -0.10. There are 3 outliers, which are sectoral groups 315, 316 and 334. Despite the sectoral differences, the estimation results obtained by considering the sectoral heterogeneity are generally in line with the previous aggregated estimation results.

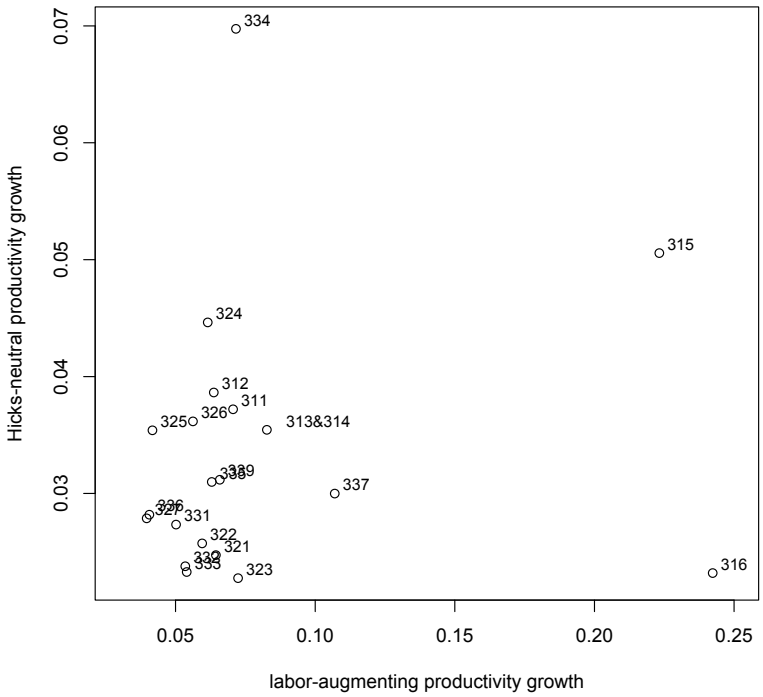


Figure 3.7: Estimation of average productivity ($\bar{\lambda}^A$ and $-\bar{\lambda}^B$)

3.5 Conclusion

In this study, I introduced a new method of estimating a CES production function with biased technical change, which extends the work of Berndt (1976), Olley and Pakes (1996), Antràs (2004), Klump et al (2007) and Leon-Ledesma et al (2010). This approach is superior to its prior counterparts in three aspects. First, it employs a more flexible production function specification; second, it is able to deal with the endogeneity problem of input variables; third, the degree of returns to scale, the elasticity of substitution and the growth rate of biased technical change can be estimated simultaneously.

The new empirical evidences presented in this chapter, show that the U.S manufacturing industries were characterized by decreasing returns to scale and non-unitary substitution elasticity (below one) technology; the bias in technical change is mainly labor-augmenting. Furthermore, the estimation results obtained by considering different windows of observation and stratified data sets, may throw some light on the questions such as the production technology evolution of last half century and the intra-industrial distortion in U.S manufacturing sectors.

3.6 Appendix

Presentation of Data set and Construction of Variables

Different sources of data

The main source of information comes from the NBER-CES Manufacturing Industry databases.[8] These data reflect essentially the Annual Survey of Manufactures (ASM) conducted by U.S. Census Bureau, which aggregates approximately 50,000 establishments to 473 six-digit NAICS manufacturing sector groups for the period 1958-2005.[9] The variables included in the database are output, employment (production/non-production), payroll, investment, capital stocks, materials and energy cost together with price deflators. The construction of this database has been discussed in the technical report of Bartlesman and Gray (1996). Malley and Muscatelli (1999) provided further detail on the definition of variables.

The variable *payroll* of the NBER data set does not include social security or other legally mandated payments, or employer payments for some fringe benefits. Therefore, the labor costs are systematically understated by this data set. In order to correct this bias, we need to include fringe benefits. To this end, additional information is required, especially the fringe benefits costs ratio, i.e, (fringe benefits/total compensation). Two sources of information can be used, i.e., the 1992-2005 ASM tables and the National Income and Product Account (NIPA) tables conducted by BEA.

The NIPA tables (especially Tables 6.2 - 6.3 and Tables 6.10 - 6.11), record the compensation of employees, wage and salary accruals, legally required social insurance, pension and insurance funds from 1948 to 2010 for 21 two-digit SIC sector groups.[10] We can use these data for covering the period of 1958 to 1991 by assuming homogeneity within sectors at the two-digit level. More disaggregated data (at four-digit SIC and six-digit NAICS level) are available in the ASM tables. For the period 1992 to 1996, we can find the value of fringe benefits recorded in SIC classification system, while for the period 1997 to 2005 the data are collected in NAICS.

Construction of capital price

The NBER database provides the total real capital stock (K), then we need to construct the rental price of capital (P_K) by using the investment price index (P_I). Consider the following formula: $P_{K,t} \equiv P_{I,t}(1+\pi_t) - \mathrm{E}_t[(1-\delta_t)P_{I,t+1}]$, where π denotes the nominal interest rate. In this study we use the 10-year U.S. treasury constant maturity rate, which comes from the Federal Reserve Bank of St. Louis. δ is the physical depreciation rate. The depreciation rate can be computed by using the classical capital accumulation

[8] The database is accessible on the website: www.nber.org/nberces/nbprod96.htm
[9] See the website: http://www.census.gov/manufacturing/asm/index.html
[10] See the website: http://www.bea.gov/national/nipaweb/Index.asp

equation, $K_t = I_t + (1 - \delta_t)K_{t-1}$ or set to be constant (in this chapter we assume that $\delta = 8\%$). Assuming that there is no expectation errors on $P_{I,t+1}$, the above formula can be simplified as: $P_{K,t} \equiv P_{I,t}(\delta_t + \pi_t)$.

Construction of fringe benefits ratio

The total fringe benefits is the employer's costs for legally required social insurance, employee pension and insurance funds.[11] The fringe benefits can be computed in two manners: the difference between the total compensation and the payroll or the sum of costs for social insurance, employee pension and insurance funds (the two methods carry out the similar results in our data). Thus, the ratio of fringe benefits to total compensation is used to magnify the labor costs of the NBER database. The main difficulty of incorporating the fringe benefits into the NBER database is that the data are recorded at different aggregation level and in different industrial classification systems before 1997. We converted the 2-digit SIC data (for the period 1958-1990) of NIPA tables and the 4-digit SIC data (for the period 1991-1996) of ASM tables in to the NAICS data, according to the concordance proposed by Census Bureau.[12]

Sources of missing values The main source of missing values is that the data on fringe benefits from the NIPA tables is only available at the 2-digit SIC level. Therefore, we assume that the fringe benefits are invariant across sectors within the 2-digit SIC industry group. The second source is that some 6-digit NAICS sectors are missing in the ASM tables for the period of 2002 to 2005. In this case, we replace the missing values by the variation rate of corresponding 5-digit NAICS sectors.[13] The third source of missing values is due to the concordance relationships between the 4-digt SIC and 6-digit NAICS classification system. Some NAICS industry groups correspond to several SIC industry groups. Thus, the fringe benefits of NAICS sector is computed as the average of fringe benefits of its SIC counterparts. In some cases, the corresponding SIC groups are not manufacturing industries. Consequently, their fringe benefits data are not available and we simply disregard these non manufacturing SIC industry groups for computing the average of fringe benefits.

Finally, we obtain a balanced panel data set that contains the output, adjusted labor costs, capital cost, investment and material input costs with price deflators for 462 NAICS manufacturing industry groups over the period 1958-2005.

[11] The ASM define the fringe benefits as the expenditures for social security tax, unemployment tax, workmen's compensation insurance, state disability insurance pension plans, stock purchase plans, union-negotiated benefits, life insurance premiums, and insurance premiums on hospital and medical plans for employees.

[12] See the website: http://www.census.gov/eos/www/naics/concordances/concordances.html

[13] The variation rate of fringe benefits at period t of 6-digit sector $= F_{t-1}^{6-digit} \cdot (1 + \frac{F_t^{5-digit}}{F_{t-1}^{5-digit}})$, where F denotes the fringe benefits rate.

Chapter 4

Fixed and Variable Cost[1]

4.1 Introduction

A long tradition going back to Viner (1931) considers that fixed costs correspond to the cost of fixed inputs.[2] However, splitting the whole set of inputs into two disjoint sets (with either fixed or variable inputs) does not provide a faithful description of many economically interesting technologies. If some variable inputs are substitutable to fixed inputs, then this sharp distinction vanishes. This chapter extends the microeconomic foundations of production analysis by allowing each input to have a fixed and a variable part.

Empirical specifications of production and cost functions are also shaped by this dichotomy between fixed and variable inputs. Some specifications consider fixed costs to be the cost of the fixed inputs. Others, like the Cobb-Douglas, the CES, and even flexible functional forms like the Translog, assume that fixed costs are nonexistent. We propose a generalization of the Translog functional form which is compatible with inputs having both a fixed and a variable part. Our empirical results support the extended Translog specification and show that the fixed cost is significant and neglecting it yield estimation biases, especially on the markup and the rate of returns to scale. Fixed costs, although not functionally dependent on the output level, are correlated with output, and should be explicitly considered to avoid these estimation biases. Our findings are compatible with the predictions of models that incorporate heterogeneous technologies (see e.g. Acemoglu and Shimer, 2000 and Cabral, 2012), in which there is a trade-off between production functions having a large fixed cost and low variable cost and those with the converse configuration.

[1] This chapter has been circulated under the title "Fixed cost, variable cost, markups and returns to scale", Chen and Koebel (2013).

[2] In the words of Viner (1931, p.26): "It will be arbitrarily assumed that all of the factors can for the short-run be sharply classified into two groups, those which are necessarily fixed in amount, and those which are freely variable. [...] The costs associated with the fixed factors will be referred to as the fixed costs".

Despite the challenging result of Baumol and Willig (1981, p.405) according to which fixed costs "do not have the welfare consequences normally attributed to barriers to entry", there is a quite large literature on fixed inputs. Fixed costs are useful for explaining coordination failure (Murphy et al., 1989) and international trade (Krugman, 1979 and Melitz, 2003). Blackorby and Schworm (1984, 1988) and Gorman (1995) have shown that fixed inputs hamper the aggregation of production (and cost) functions, whereas a fixed cost does not represent an aggregation problem. Fixed costs are also considered in general equilibrium theory with imperfect competition, see for instance Dehez et al. (2003). Contributions in the field of industrial organization on the reasons and consequences of fixed (and sunk) cost, are so numerous that we cannot survey them here. Berry and Reiss (2007) discuss some important issues on identification and heterogeneity of fixed costs. Differences between fixed and sunk cost are commented by Wang and Yang (2001, 2004) and Sutton (2007).

The objective of this chapter is to characterize and estimate both fixed and variable components of the cost function, to investigate their heterogeneity over firms and study how fixed costs affect their behavior in terms of price setting and returns to scale. Microeconomic textbooks present alternative characterizations of fixed costs. We follow Baumol and Willig (1981, p.406) and consider the long run fixed cost as the magnitude of the total long run cost function when the production level tends to zero. This chapter derives the production technology which generates the fixed cost, an issue which is usually neglected when dealing with fixed cost. It is well known (see Mas-Colell et al., 1995, p.135) that fictitious inputs can be used for imposing constant returns to scale on arbitrary technologies. This chapter shows that the fixed cost of production can be represented as the cost of fictitious (unobserved) inputs. We first characterize the production technology which generates the traditional fixed cost and show that it is quite restrictive and given by $y = F(x_v + x_f)$ where x_v denotes the vector of variable inputs and x_f the fixed inputs. As total input x can always be additively split into two categories, the structure F may be considered as perfectly general. However, two physically similar inputs may be technologically different and we propose to extend the production function to $y = G(x_v, x_f)$. This extended production technology generates a fixed cost which is not equal to the cost of inputs x_f, and identification of fixed inputs is no longer possible. However, the amount of inputs which allows to initiate production is well identified.

Our theoretical contribution also requires extending the econometric toolbox for estimating cost functions. First, usual cost function specifications are not compatible with a flexible specification of the fixed cost. For approximating a cost function with a fixed cost component, we have to go beyond (locally) flexible cost functions, and develop a cost specification which is a valid approximation at two points: around the actual point of production and around the breakup point which allows a firm to start production.

Second, as the inputs x_v and x_f cannot be observed, we have to amend the traditional estimation method by introducing unobserved and correlated heterogeneity in the fixed and variable cost specification. We extend Swamy's (1970) random coefficient estimator to our nonlinear setup. The empirical part of this chapter uses panel data for U.S. manufacturing sectors in order to estimate the size and the type of fixed cost as well as their implications in terms of markup pricing, returns to scale and technical change.

In Sections 4.2 and 4.3 we explore two definitions of fixed costs and their microeconomic foundations. Sections 4.4, 4.5, and 4.6 discuss econometric issues related to fixed costs: biases when they are neglected, specification issues, and unobserved heterogeneity. Section 4.7 reports the empirical results, obtained for 462 U.S. manufacturing industries observed over the years 1958 to 2005.

4.2 Defining fixed costs and fixed inputs

The definition of fixed costs is central in economics and is briefly discussed in most introductory microeconomic textbooks.[3] One difficulty with most definitions is that they do not highlight the relationship between the fixed cost and the fixed inputs. Are fixed inputs physically fixed? Do fixed inputs correspond to non-optimal choices? This section shows that it is not necessarily the case: a fixed cost can arise in a context where all inputs are optimally adjusted.

Most economists agree that the fixed cost u corresponds to the part of the cost which does not vary with the level of production:

$$c(w,y) = u(w) + v(w,y), \tag{4.1}$$

where w denotes the input prices and y the output level. Function v corresponds to the variable cost of production and satisfies $v(w,0) = 0$. Any cost function can uniquely be written in this way by defining:

$$\begin{aligned} u(w) &\equiv c(w,0) \\ v(w,y) &\equiv c(w,y) - c(w,0). \end{aligned} \tag{4.2}$$

We will comment the following alternative definitions for the fixed cost and fixed inputs.

Definitions 4.1. *For an active firm, the* fixed cost *is*

a) the accounting cost of the inputs which are physically fixed.

b) the cost of the inputs required for producing an arbitrarily small amount of output.

[3] It seems somewhat surprising, however, that the New Palgrave dictionary of economics has no entry for the term "fixed cost". The term is also not commented in Diewert's (2008) contribution on cost functions.

Definition 4.1 does not require (at this stage) that the level of the fixed cost is optimal (so it does not necessarily correspond to the minimal value of the accounting cost). In Definition 4.1b the inputs required for initiating production could be physically fixed but it is not necessary the case. Since the cost function is related to input demands x° by the accounting relationship $c(w,y) = w^{\top}x^{\circ}(w,y)$ for any $y \geq 0$, we obtain the level of fixed cost compatible with Definition 4.1b as:

$$u(w) = \lim_{y \to 0^{+}} c(w,y) = \lim_{y \to 0^{+}} w^{\top}x^{\circ}(w,y).$$

This shows that Definition 4.1b implies that the fixed cost does not change with the production level, but can change with w. Whereas it is straightforward to define *variable inputs* as inputs whose level can be adjusted to minimize their accounting cost and can *possibly* be set to zero, the definition of fixed inputs is more involved, as they are not *necessarily* optimal, nor can they *necessarily* be set to zero.

We show that the fixed cost $u(w)$ does not necessarily correspond to the cost of the fixed inputs, but that it also includes a part of the cost of variable inputs when they are sufficiently complementary to the fixed inputs. For instance, if capital is physically fixed and energy is fully variable, but capital cannot be run without say 1000 KWh of energy, then the part of the energy input which is necessary to run the fixed capital input becomes fixed. It is the production technology which determines whether inputs are variable or fixed and which part of each input is fixed or variable. This remark has important implications for the specification of fixed and variable cost functions and these have been largely ignored in the literature.

4.3 A microeconomic framework for fixed costs

The main result of this section characterizes an extended production function able to describe fixed inputs in a more general way than the existing literature. A shortcoming of the traditional restricted cost function (see Subsection 4.3.1), is that it relies on a partition of all inputs into two *disjoint* categories: variable and fixed inputs. Actually, similar inputs can be used for different types of production activities. Engineers, for instance, can be allocated to production or to research and development activities. While engineer's production increases the *current* output level, it is not the case when they are allocated to research and development, which withdraws them from production (like in Aghion and Howitt, 1992, for instance). Similarly, computers can be used either for logistics, production management or accounting, activities which do not have the same impact in terms of production and cost. Before presenting the extended production and cost function, we shortly overview traditional production analysis.

4.3.1 On the limitations of traditional production analysis

For modeling fixed inputs, production analysis relies on a partition of the input vector x into two *disjoint* categories: those which can be adjusted (variable inputs, denoted $\widetilde{x}$) and those which are fixed or quasi-fixed ($\overline{x}$):[4]

$$x = \begin{pmatrix} \widetilde{x} \\ \overline{x} \end{pmatrix} \geq 0. \tag{4.3}$$

The corresponding input prices are denoted by $\left(\widetilde{w}^\top, \overline{w}^\top\right)^\top$. The output level is given by $y = F(x)$ where $F : \mathbb{R}^J \to \mathbb{R}$ denotes the production function which is increasing in x. The *restricted* variable cost function is defined as:

$$V_r(\widetilde{w}, \overline{x}, y) = \min_{\widetilde{x} \geq 0} \left\{ \widetilde{w}^\top \widetilde{x} : F(\widetilde{x}, \overline{x}) \geq y \right\}.$$

The properties of the restricted cost functions have been investigated by Lau (1976) and Browning (1983). For empirical implementations see e.g., Caves et al. (1981), Pindyck and Rotemberg (1983) and Morrison (1988). The total restricted cost function is given by:

$$V_r(\widetilde{w}, \overline{x}, y) + \overline{w}^\top \overline{x}, \tag{4.4}$$

where the last term denotes the fixed cost. In the long-run, all the fixed inputs can be adjusted at their optimal level and this defines the long-run or *unrestricted* cost function:

$$c(w, y) = \min_{\overline{x} \geq 0} \left\{ V_r(\widetilde{w}, \overline{x}, y) + \overline{w}^\top \overline{x} \right\} = \widetilde{c}(w, y) + \overline{c}(w, y), \tag{4.5}$$

where $\widetilde{c}(w, y) = V_r\left(\widetilde{w}, \overline{x}^*(w, y), y\right)$ represents the long run variable cost, and $\overline{c}(w, y) = \overline{w}^\top \overline{x}^*(w, y)$ is the long run fixed cost. Function $\overline{x}^*$ denotes the optimal level of fixed inputs, which, without further restrictions on V_r, depends on the production level. As a consequence, this approach violates (in the long run) both definitions given in Definition 4.1. More than that, in the long run it is not possible to identify $\widetilde{c}$ separately from $\overline{c}$, unless we make strong *a priori* assumptions on which inputs are fixed in the short run. A further drawback of technology F appears when we impose that V_r be a variable cost function, namely $V_r(\widetilde{w}, \overline{x}, 0) = 0$. This restriction implies that there are no fixed cost in the long run: $\overline{x}^*(w, 0) = 0$ (unless we impose a positive lower bound to $\overline{x}$).

So, according to traditional production analysis, the only justification for fixed cost is that physically fixed inputs cannot be optimally adjusted (either for technical reasons or for lack of rationality). This view excludes a variety of interesting situations in which fixed and variable inputs are imperfect substitutes and play different roles in production.

[4]Here, the notation $x > 0$ means that all J components $x_j > 0$. In contrast $x \geq 0$ means that $x_j \geq 0$ for all j.

4.3.2 Another view of the traditional production function

Instead of partitioning x into two disjoint types of inputs, let us assume that each input comprises a part which can be adjusted and a part which is fixed (in a sense that is clarified in Definition 4.2 below):

$$x = x_v + x_f, \tag{4.6}$$

with $x, x_v, x_f \in \mathbb{R}^J_+$. This generalizes (4.3) which is obtained as a special case when $x_v = \left(\widetilde{x}^\top, 0^\top\right)^\top$ and $x_f = \left(0^\top, \overline{x}^\top\right)^\top$. This subsection shows that the variable and fixed cost functions used in production analysis is generated from an additive production function:

$$y = F\left(x_v + x_f\right), \tag{4.7}$$

which requires perfect substitutability between x_v and x_f.

As our purpose is to describe the production possibilities for a production level close to zero (in order to be consistent with Definition 4.1b), we define the input requirement set as follow.

Definition 4.2. *In terms of the traditional production function, the fixed cost is the cost associated to inputs belonging to the* input requirement set X_F *defined as:*

$$X_F \equiv \lim_{\varepsilon \to 0^+} \left\{z \geq 0 : F\left(z\right) = \varepsilon\right\}.$$

Definition 4.2 requires that the limiting isoquant X_F exists. Definition 4.2 is useful to characterize the fixed cost in terms of the production function F: it is easy to show that a fixed cost occurs if the set X_F does not include the point $x = 0$.[5] In order to be compatible with Definition 4.1b, we consider in Definition 4.2 the isoquant corresponding to the production level $\varepsilon > 0$, instead of $\varepsilon = 0$, because with most production functions compatible with a fixed cost, the condition $F\left(x\right) = 0$ characterizes a thick isoquant, in the sense that, if it is possible to produce nothing with something ($\exists x > 0 : F\left(x\right) = 0$), then it is also possible to produce nothing with even less (there exist $x' < x$ such that $F\left(x'\right) = 0$). So, only the upper frontier of the set $\left\{z \geq 0 : F\left(z\right) = 0\right\}$ is interesting for identifying a fixed cost. Let us investigate the implications of this additive structure in

[5] For the purpose of exposition we assume that the minimum value of y included in the range of F is zero, but we could easily generalize to any other value.

terms of the restricted variable and total cost functions:

$$\begin{aligned} v_r(w, x_f, y) &= \min_{x_v \geq 0} \left\{ w^\top x_v : F(x_v + x_f) \geq y \right\}; \\ c_r(w, x_f, y) &= v_r(w, x_f, y) + w^\top x_f. \end{aligned}$$

The restriction $x_v \geq 0$ is important here, because it can be optimal to use no variable inputs at all for some levels of x_f.

Proposition 4.1. *Let $x_f \in X_F \neq \varnothing$ and $x_v \geq 0$. Then $c_r(w, x_f, 0) = w^\top x_f \geq 0$, $v_r(w, x_f, 0) = 0$. The restricted cost function c_r and the cost minimizing variable inputs x_v^* satisfy either*

(i) *for $x_v^* > 0$ and $y > 0$,*

$$c_r(w, x_f, y) = C(w, y) > w^\top x_f \tag{4.8}$$

and $x_v^*(w, x_f, y) = X_v^*(w, y) > 0$ or

(ii) *$x_{v,j}^* = 0$ for some j, and*

$$c_r(w, x_f, y) = V_r(\widetilde{w}, \overline{x}, y) + \overline{w}^\top \overline{x}, \tag{4.9}$$

where $\overline{x}$ is a subvector of x_f and $\widetilde{w}$ corresponds to the price subvector of $w = \left(\widetilde{w}^\top, \overline{w}^\top\right)^\top$ corresponding to $x_{v,j}^* > 0$.

The proof of this result is given in the Appendix. Proposition 4.1 states that the variable cost is zero when production vanishes. This result is driven by the additive structure of F which ensures that if there exists a point x such that $F(x) = 0$, then x_v can be set to zero in the additive decomposition $x = x_v + x_f$. In the case of Proposition 4.1(i), the production function F yields a cost function which is independent of the level of fixed input, and which is compatible with both Definition 4.1a and 1b. A frustrating consequence of Proposition 4.1(i) is that fixed inputs can be seen as if they were set at their optimal level, as:

$$\frac{\partial c_r}{\partial x_f}(w, x_f, y) = 0 \Leftrightarrow -\frac{\partial v_r}{\partial x_f}(w, x_f, y) = w.$$

It is the perfect substitutability between the variable and the fixed inputs which is driving this result. Any mistake in adjusting x_f can be perfectly compensated by setting x_v optimally. In summary, technology F is not really suitable for modeling fixed inputs, as it lacks generality. Proposition 4.1(ii) gives the general formulation of the cost function corresponding to F when corner solutions for the variable inputs are

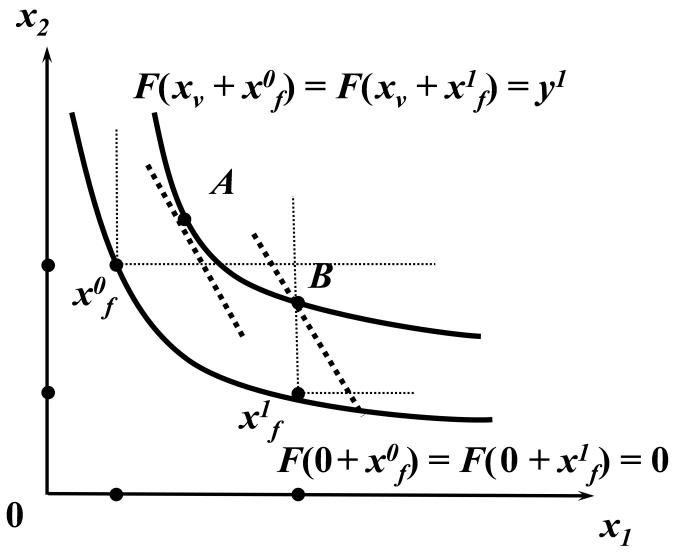


Figure 4.1: Isoquants for $F(x_v + x_f)$

allowed. The structure of the cost function (4.9) is the same as in (4.4) and is common in traditional production analysis (see Chambers, 1988, for instance). So we conclude this section by noting that production function F with an additive structure between x_v and x_f is behind the traditional theory of fixed and variable costs. This additive structure is restrictive and hides important features of production theory.[6]

Figure 4.1 illustrates Proposition 4.1. Endowed with a fixed input vector x_f^0, the variable inputs available to the firm and satisfying $x_v \geq 0$ are located in the north-east quad of x_f^0. At given input prices w, the firm minimizes its variable cost of producing y^1 at the interior point A. At this point, according to Proposition 4.1(i), the cost function is given by $C(w, y)$. With another level of fixed inputs, however, the available set of variable inputs will be different. With x_f^1, the minimal variable cost for producing y^1 is achieved at B, on the boundary of the set $\left\{x_v \geq 0 : F\left(x_v + x_f^1\right) \geq y^1\right\}$. At point B we have $x_{v1}^* = 0$ and the optimal level of x_{2v}^* is restricted by the level of x_f^1.

4.3.3 An extended production function

Whereas from the accounting viewpoint both types of inputs x_v and x_f are similar (the cost of a unit of the j^{th} fixed and flexible input is w_j), technologically they should not be restricted to play similar roles as it is the case with $F(x_v + x_f)$. We now define

[6] One restriction is that

$$\frac{\partial^2 c_r}{\partial \widetilde{w}_j \partial \overline{w}_k}(w, x_f, y) = 0.$$

For given x_f, y, there is no substitutability between inputs j and k. This is too restrictive because, even for given x_f, inputs j and k can be substituted for each other because they have a fixed and a variable component.

an extended production function G as $y = G(x_v, x_f)$ where $G : \mathbb{R}^J_+ \times \mathbb{R}^J_+ \to \mathbb{R}_+$. For simplicity, we assume that G is single valued, continuously differentiable, increasing in its arguments and that $G(0, x_f) = 0$. In this context, the restricted variable cost function now becomes:

$$v_r(w, x_f, y) = \min_{x_v} \left\{ w^\top x_v : G(x_v, x_f) \geq y \right\}. \tag{4.10}$$

Now, a given input, say capital, can appears twice in (4.10): once in vector x_v and once in x_f; their marginal productivities can be different. This overlapping structure is similar to the one considered by Blundell and Robin (2000) in consumer analysis. In contrast to their approach, we do not impose that x_v is separable from x_f (a structure which they call latent separability).

Leontief's (1947) aggregation theorem highlights the restrictions which are implicit in production function F. The number $2J$ of inputs x_v and x_f which appear in G can be reduced to the J aggregate inputs $x_v + x_f$ iff we have:

$$\frac{\partial G}{\partial x_{v_i}} = \frac{\partial G}{\partial x_{f_i}}, \quad \forall i = 1, \ldots, J.$$

We do not assume in the sequel that these restrictions necessarily apply to G.

One difficulty with (4.10), is that if v_r is defined for any arbitrary levels of x_f, we can switch the notation from x_f to x_v and rewrite $v_r(w, x_v, y)$. So, in order to be able to identify x_f as the fixed inputs, we need to put more structure on v_r, and we do this by introducing restrictions derived from the definition of the fixed cost and inputs.

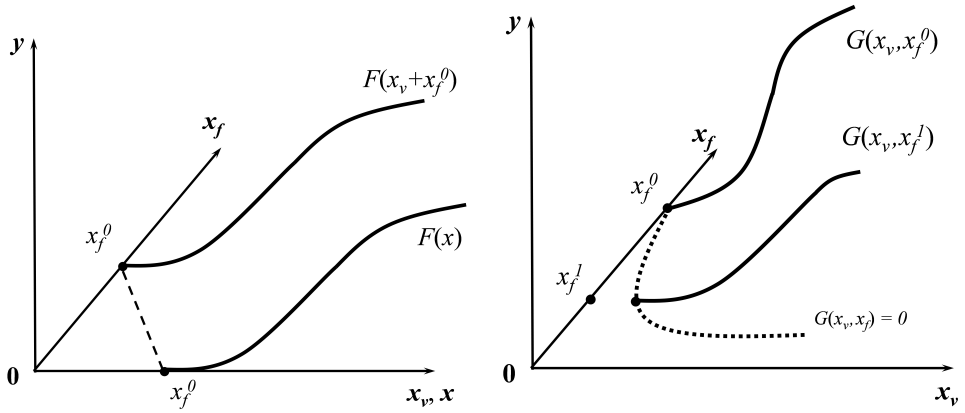
Definition 4.3. *In terms of the extended production function G, the fixed cost is the cost associated to inputs belonging to the* fixed input requirement set X_G *defined as:*

$$X_G \equiv \lim_{\varepsilon \to 0^+} \{ z \geq 0 : G(0, z) = \varepsilon \}. \tag{4.11}$$

Definition 4.3 defines the set of all fixed input combinations required for starting production. This definition is more general than Definition 4.2, because it does not assume that fixed and variable inputs are perfectly substitutable. As for X_F, we impose that $x_v = 0$ belongs to the fixed input requirement set X_G, but get rid of additivity. The next result is a straightforward extension to technology G of those available for technology F.[7]

Proposition 4.2. *If $x_f \in X_G$ then,*

[7] We only give the properties which are the more interesting for our purpose, see Lau (1976) and Browning (1983) for other properties.



(a) Technology F (b) Technology G

Figure 4.2: Fixed and variable inputs and production possibilities

(i) $v_r(w, x_f, 0) = 0$, $v_r(w, x_f, y) > 0$ *for any* $y > 0$

(ii) v_r *is increasing in* y

(iii) v_r *is decreasing in* x_f.

Proposition 4.2 means that the restricted variable cost function v_r satisfies the properties of a variable cost function: it vanishes for arbitrarily small production levels. As a consequence, the restricted fixed cost is given by:

$$u_r(w, x_f) \equiv \lim_{y \to 0^+} c_r(w, x_f, y) = w^\top x_f,$$

and total restricted cost satisfies

$$c_r(w, x_f, y) = u_r(w, x_f) + v_r(w, x_f, y). \tag{4.12}$$

Both production technologies F and G are represented on Figure 4.2 in the case where a single input is decomposed into a fixed and a variable component. Figure 4.2a illustrates how the introduction of a fixed input x_f satisfying $F\left(x_f^0\right) = 0$ and the reparameterization $x \equiv x_v + x_f$ yield the technology $F\left(x_v + x_f^0\right)$. On Figure 4.2b, the isoquant corresponding to the startup production level $G(x_v, x_f) = \varepsilon$ is not a straight line, which opens the possibility to choose a fixed input different from x_f^0 as an admissible value for starting production. Input x_f^1 for instance, allows to start production with production function $G\left(x_v, x_f^1\right) \neq G\left(x_v, x_f^0\right)$, provided that x_v is sufficiently high for compensating the decline from x_f^0 to x_f^1.

We illustrate the usefulness of technology G with an example which also illustrates the claims of Proposition 4.2.

Example. The technology $G : \mathbb{R}^2_+ \to \mathbb{R}_+$ is given by:

$$y = G(x_v, x_f) = (x_v + \beta x_f) x_f^{\alpha} - \gamma$$

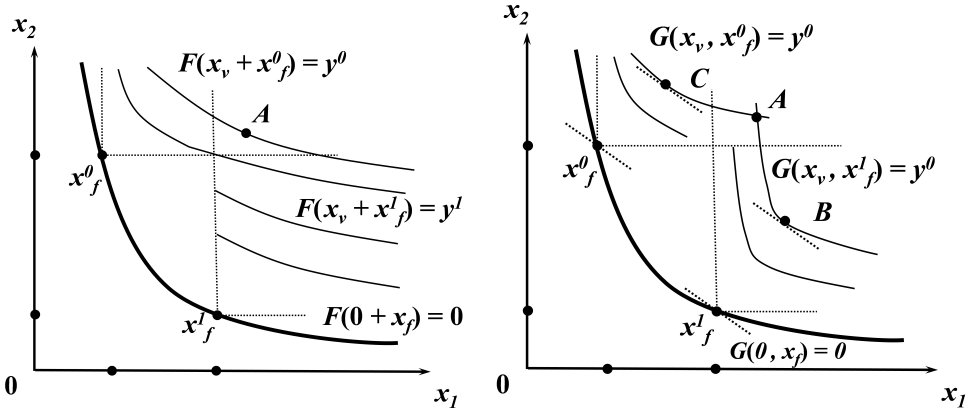
for $y \geq 0$. Here $x_f \in X_G \Leftrightarrow x_f = (\gamma/\beta)^{1/(\alpha+1)}$. This yields the restricted variable cost function:

$$v_r(w, x_f, y) = w x_v^*(w, x_f, y) = w(y + \gamma) x_f^{-\alpha} - w\beta x_f = w \frac{y}{x_f^{\alpha}},$$

which satisfies $v_r(w, x_f, 0) = 0$ for $x_f = (\gamma/\beta)^{1/(\alpha+1)}$. The restricted fixed cost function is $u_r(w, x_f) = w x_f = w(\gamma/\beta)^{1/(\alpha+1)}$. For $\alpha = 0$ and $\beta = 1$ we obtain the traditional production function as a special case. This example also illustrates that in both cases of exogenous (physically fixed) and endogenous input x_f, there is no conflict between Definition 4.1a and 4.1b.

The structure of the isoquants of F and G is represented in Figure 4.3 for $J = 2$, in the (x_1, x_2)-plane (with $x_1 = x_{v1} + x_{f1}$). In Figure 4.3a the slopes of the isoquants corresponding to F only depend upon total input use $x = x_v + x_f$ and not upon the share of the fixed inputs x_f in the composite input x. At point A for instance, it is possible to produce y^0 using fixed input x_f^0 or x_f^1. Only the total input quantity matters and since $x_f^0 + x_v^0 = x_f^1 + x_v^1$ at point A, the choice of the fixed input is irrelevant. Note that, contrary to Figure 4.2, the isoquants do not necessarily cross the axes on Figure 4.3, because axes now report total input levels for two different inputs, and not just how a given input is split into variable and fixed amounts.

Figure 4.3b represents in the (x_1, x_2)-plane the isoquants for technology G and two different fixed input vectors x_f^0 and x_f^1. With technology G, the choice of the level of fixed inputs determines the substitution possibilities between the variable inputs. Although we have not introduced any distinction between *ex-ante* and *ex-post* technologies in our model, Figure 4.3 resembles those typically obtained with putty-putty (or putty-clay or clay-clay) technologies (see e.g., Fuss, 1977). The similarity is due to the fact that we split x into two (fixed and variable) non-additive components. With technology G the choice of a particular fixed input level x_f coincides with a choice of a particular production technology and a specific substitution pattern between variable inputs. On Figure 4.3b, the isoquant corresponding to x_f^0 characterizes inputs which can easily be substituted the one for the other, whereas for x_f^1 substitution becomes more difficult. Note that for a given output level, the isoquants for G corresponding to the fixed input level x_f^0 can cross those obtained for x_f^1. For instance, at point A the production



(a) Isoquants for F (b) Isoquants for G

Figure 4.3: Fixed and variable inputs and substitution possibilities

level y^0 can be produced using two types of technologies, each one exhibiting a specific substitution pattern.

Figure 4.3b also illustrates that if fixed inputs are neglected, production function G is not necessarily quasi-concave in x (at point A). Moreover, optimal choices for input bundles can be located in the zone violating quasi-concavity in x and so the cost function will not necessarily be concave in w. In the context of fixed cost, imposing simultaneously concavity in w and $x_f = 0$ on the cost function may end up with worse estimates than extending the cost function to be compatible with the occurrence of fixed cost (see Lau, 1978, and Diewert and Wales, 1987, for seminal contributions on concavity enforcement).

The next difficulty we have to deal with is related to the fact that the level of fixed inputs can be either exogenous or endogenous. Figure 4.3b depicts at point C a situation at which the variable inputs are optimal given the levels of fixed inputs x_f^0 and production level y^0, however, if x_f could be chosen, the firm would set them to x_f^1 and produce y^0 at point B. It is important to note that isoquant and isocost line are not necessarily tangent at the optimum level x_f^1 for $x_v = 0$.

Whereas variable inputs can by definition be adjusted for minimizing costs, the fixed inputs are not necessarily set at their optimal level. We say that a fixed input x_{fj} is *exogenous* when its actual level is not optimal in the sense that the equality between its shadow value and market price is violated:

$$-\frac{\partial v_r}{\partial x_{fj}}(w, x_f, y) \neq w_j, \tag{4.13}$$

for the observed values of (w, x_f, y) and $x_{fj} > 0$. The extended framework based on $G(x_v, x_f)$ is useful as it allows to split the input x into a part x_v that is efficiently

allocated, and a part x_f which is not necessarily so.[8]

In the long run, fixed inputs can be determined endogenously by the firm, and they may in some case be set to zero. Such a corner solution occurs at $x_f = 0$ if $0 \in X_G$ and:

$$0 \leq v_r(w, 0, y) \equiv \min_{x_v \geq 0} \left\{ w^\top x_v : y \leq G(x_v, 0) \right\} < v_r(w, x_f, y) + w^\top x_f,$$

for any $x_f > 0$. Equivalently, the choice $x_f^* = 0$ is (locally) optimal if at point $(w, 0, y)$ the increase in fixed cost is not compensated by a greater reduction of the variable cost:

$$w_j + \frac{\partial v_r}{\partial x_{fj}}(w, 0, y) > 0.$$

Then it is optimal to adopt a production structure without any fixed input. In many cases however, an inner solution for x_f^* exists. It is characterized by the equality between the shadow value of the fixed input and its market price:

$$-\frac{\partial v_r}{\partial x_{fj}}\left(w, x_f^*, y\right) = w_j. \tag{4.14}$$

Example (continuation). For $v_r(w, x_f, y) = w(y + \gamma) x_f^{-\alpha} - w\beta x_f$, we find that (assuming $\alpha > 0$ and $\beta < 1$):

$$x_f^*(w, y) = \left(\frac{\alpha}{1 - \beta}(y + \gamma)\right)^{\frac{1}{1+\alpha}},$$

which varies with the level of output. In the traditional case: $\alpha = 0$ and $\beta = 1$, the restricted variable cost function becomes $v_r(w, x_f, y) = wy$ and we obtain a corner solution $x_f^* = 0$, conformably to Section 4.3.1. The long-run variable cost function becomes:

$$v_r\left(w, x_f^*, y\right) = w(y + \gamma)\left(\frac{\alpha}{1 - \beta}(y + \gamma)\right)^{\frac{-\alpha}{1+\alpha}} - w\beta\left(\frac{\alpha}{1 - \beta}(y + \gamma)\right)^{\frac{1}{1+\alpha}},$$

and this does not necessarily vanish anymore for a production level going to zero:

$$v_r\left(w, x_f^*(w, 0), 0\right) = w\gamma\left(\frac{\alpha}{1 - \beta}\gamma\right)^{\frac{-\alpha}{1+\alpha}} - w\beta\left(\frac{\alpha}{1 - \beta}\gamma\right)^{\frac{1}{1+\alpha}}.$$

The example above illustrates the fundamental identification problem occurring when inputs are optimally adjusted: the fixed cost generally differs from the cost of the fixed inputs. Indeed, after normalizing the variable and fixed cost function accord-

[8] Common explanations for why the level of the fixed inputs is not optimal are related to (i) technological constraints, (ii) indivisibilities of the fixed inputs, (iii) allocative inefficiencies and (iv) inter-temporal dependences.

ing to (4.2), we obtain the fixed cost:

$$u(w) = w^{\top} x_v^* \left(w, x_f^*(w,0), 0\right) + w^{\top} x_f^*(w,0).$$

When fixed and variable inputs can be imperfectly substituted for each other, the optimal amount of fixed input depends upon w and $x_f^*(w,0)$ is not necessarily included in the input requirement set X_G. This means that the level of fixed input cannot be determined ex-ante using only the definition of X_G. When x_f can be adjusted, it is no longer possible to separately identify x_f and x_v. Fortunately, Definition 4.1b of the fixed cost is fully compatible with this situation, but Definition 4.1a is violated: $u(w) \neq w^{\top} x_f^*(w,0)$. Briefly, an input cannot be said to be fixed or variable *prima facie*, using only physical properties of the inputs. It is the technology which in last instance determines whether a given input is fixed or variable. This explains why Definition 4.1b which relies on the technology provide the more general definition of the fixed cost. Few technologies allow to obtain an optimal level of x_f^* independent of y. We characterize them below.

Proposition 4.3. *Assume that the technology G is increasing and quasi-concave in x_v, and that $x_v^* > 0$ at the optimum. Let $K : \mathbb{R}_+^J \to \mathbb{R}_+^J$ and $F : \mathbb{R}_+^J \to \mathbb{R}_+$ both be increasing functions.*

(i) *The restricted cost function is given by:*

$$c_r(w, x_f, y) = u_r(w, x_f) + v(w, y), \tag{4.15}$$

with $v(w,0) = 0$ if and only if the production function is given by:

$$G(x_v, x_f) = F(x_v + K(x_f)). \tag{4.16}$$

(ii) *The optimal level of x_f is independent of y if and only if the restricted cost function is (4.15) or the production function is (4.16).*

Proposition 4.3 characterizes the cost and production functions which generate a fixed cost. Requirement (4.16) is less stringent than separability of G in x_f because it does not impose that $K(x_f)$ be a unique aggregate fixed input. Here, the vector valued function K comprises J aggregates for the fixed inputs. Proposition 4.3 also aggregates additively some fixed and variable inputs together since F depends upon $x_v + K(x_f)$. As can be seen by comparing (4.16) and $F(x_v + x_f)$, the former is also more general than the traditional production function F for which fixed and variable inputs are perfect substitutes. Figure 4.4 provides an illustration in the two inputs case ($J = 1$). It shows that x_f does

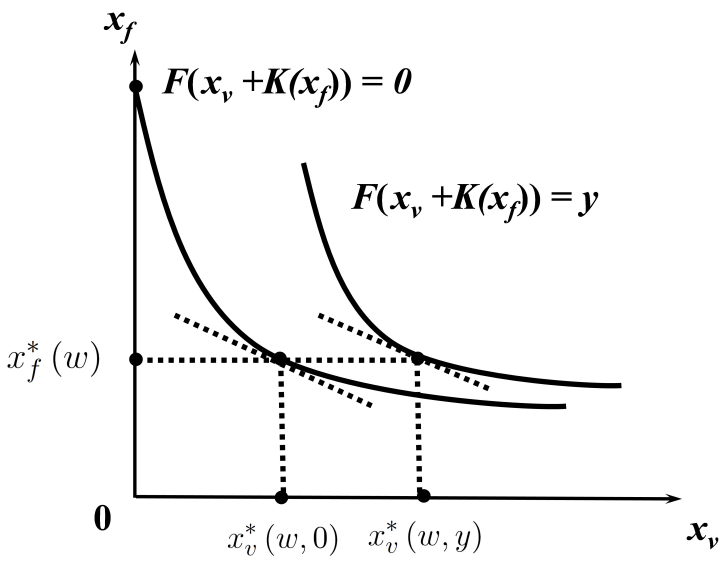


Figure 4.4: Fixed and variable inputs decomposition

not vary with y, contrary to x_v^*.[9] Figure 4.4 gives the decomposition of variable input x_v into a fully variable component $x_v^*(w,y) - x_v^*(w,0)$ which can be set to zero when there is no production, and $x_v^*(w,0)$ which has to be used for starting production. It also shows how technology $F(x_v + K(x_f))$ differs from $F(x_v + x_f)$. With $F(x_v + K(x_f))$ there is perfect substitutability between the components of x_v and $K(x_f)$, but not between x_v and x_f. For a given input x_i, the slope $(\partial F/\partial x_{vi}) / (\partial F/\partial x_{fi})$ of the isoquant (Figure 4.4) is not restricted to be equal to -1 out of the optimum. Moreover, for two different inputs, x_h and x_i, the slope $(\partial F/\partial x_{fh}) / (\partial F/\partial x_{fi})$ of the isoquant is not restricted to be equal to $(\partial F/\partial x_{vh}) / (\partial F/\partial x_{vi})$ out of the optimum. Fixed inputs can be substituted according to a different pattern than variable inputs.

We conclude this section by emphasizing that, even though a separate identification of x_f and x_v is not possible without additional restrictions, it is possible to identify uniquely $x_f^*(w,0) + x_v^*(w,0)$ as well as the level of the fixed cost. If we assume that decomposition (4.1) is not unique, then there exist $\widetilde{u} \neq u$ and $\widetilde{v} \neq v$ such that:

$$c(w,y) = \widetilde{u}(w) + \widetilde{v}(w,y),$$

with $\widetilde{v}(w,0) = v(w,0) = 0$. However, the equality:

$$u(w) + v(w,y) = \widetilde{u}(w) + \widetilde{v}(w,y)$$

[9] We also see why the corner solution $x_v^* = 0$ has to be excluded, because at this point the level of x_f^* can vary with y.

is satisfied for any (w, y) iff $u(w) = \widetilde{u}(w)$ (obtained for $y = 0$) and $v(w, y) = \widetilde{v}(w, y)$, and this proves unicity. It is also interesting to note, that although fixed cost cannot be observed, because the situation in which firms produce an output level close to zero is hypothetical, the level of fixed cost is well identified empirically and can be estimated.

4.4 Some consequences of neglecting fixed costs

This section discusses three drawbacks arising when fixed inputs are neglected. A first problem of disregarding x_f is the oversimplification of various economic relationships, in particular the relationship between fixed inputs and pricing behavior. Let $p = P(y, z)$ denote the inverse output demand which depends on exogenous macroeconomic parameters z and the firm's own production level. With market power, the firms' optimum is characterized by:

$$\frac{\partial v_r}{\partial y}(w, x_f, y) = p\left(1 + \frac{\partial P}{\partial y}\frac{y}{P}\right). \tag{4.17}$$

This equation and the discussion above shows that a fixed input x_f has an impact on the marginal cost function unless c_r has the specific structure given in (4.15). It also implies that there is a relationship between the fixed input and the markup $\eta \equiv \partial \ln P / \partial \ln y$, via the marginal cost.

Neglecting the fixed cost is a source of bias. By Shephard's lemma, we have:

$$x^*(w, y) = \frac{\partial u}{\partial w}(w) + \frac{\partial v}{\partial w}(w, y).$$

If the fixed cost is neglected, then it enters the residual term which will be correlated with w, and may bias the estimates.

From a theoretical viewpoint, neglecting the fixed cost by setting u (or u_r) equal to zero may lead to underestimation of returns to scale. In order to show this point, we consider the long-run case and assume that the cost function is convex in y. By convexity we have:

$$\begin{aligned} & c(w, 0) \geq c(w, y) + \frac{\partial c}{\partial y}(w, y)(0 - y) \\ \Rightarrow \quad & \frac{\partial c}{\partial y}(w, y)\frac{y}{c(w, y)} \geq 1 - \frac{c(w, 0)}{c(w, y)}. \end{aligned}$$

As the return to scale is the inverse of the cost elasticity with respect to the output, imposing zero fixed cost implies imposing decreasing return to scale. The equation above also shows that for given level of costs and outputs, neglecting the fixed cost leads to an overestimation of the marginal cost, which will also cause an underestimation of the markup. As $\partial c / \partial y(w, y) = w^\top \partial x^* / \partial y(w, y)$, overestimating the marginal cost often coincides with the overestimation of the input demand sensitivity to output

variations. In addition, from an empirical viewpoint, setting the fixed cost equal to zero introduces an omitted-variable bias in the estimation of technology parameters. In the following sections, we discuss the empirical issues raised by the estimation of the fixed cost, including suitable functional forms for cost functions, and the treatment of cost heterogeneity with unobserved levels of x_f.

4.5 On flexible functional forms

In the 1970's and 1980's, several researchers proposed new parametric specifications for the production technology, and introduced so-called flexible functional forms, which are able to approximate locally an arbitrary cost function. These functional forms, still widely used in production analysis, are not adequate for modeling fixed costs: either they completely exclude fixed costs, or specify them in an inflexible way. The variable t is now introduced for denoting technical change.

In their seminal paper, Diewert and Wales (1987) have introduced several cost functions, many of which can be written as:

$$C^{DW}(w,y,t) = a_w^\top w + \left(\alpha_w^\top w\right) a_t t + V^{DW}(w,y,t), \tag{4.18}$$

with $V^{DW}(w,0,t) = 0$. This identifies the fixed cost as $U^{DW}(w,t) = a_w^\top w + \left(\alpha_w^\top w\right) a_t t$, where a_w, α_w, a_t denote technological parameters. So, the fixed cost function is linear in w and t and is not a flexible specification (in the sense of Diewert and Wales, 1987). The same can be shown for the variable cost specification V^{DW}.

Let us now consider the Translog functional form (Christensen et al., 1971) with technology parameters given by β:

$$\begin{aligned} C^{TL}(w,y,t) \;=\; & \exp(\beta_0 + \beta_w^\top \ln w + \beta_y \ln y + \beta_t t \\ & + \frac{1}{2}\ln w^\top B_{ww} \ln w + \ln w^\top B_{wy} \ln y + \ln w^\top B_{wt} t \\ & + \frac{1}{2}\beta_{yy}(\ln y)^2 + \beta_{yt} t \ln y + \frac{1}{2}\beta_{tt} t^2), \end{aligned} \tag{4.19}$$

where the notation is as in Koebel et al. (2003). One of the main drawbacks of the Translog functional form is that it is not suitable for modeling fixed cost.

Proposition 4.4. *The Translog functional form implies a fixed cost that is either zero or infinite (in which case C^{TL} is decreasing in y for some values of y).*

This result shows that the Translog cost function is badly behaved in some regions, and especially when production is close to zero, which defines the fixed cost of production. This proposition illustrates that the Translog is only able to approximate locally

an unknown cost function, but not globally, and justifies the specification of alternative functional forms for the purpose of estimating a fixed cost. Proposition 4.4 points out a paradox: although the Translog specification is flexible (Diewert and Wales, 1987, Theorem 1), it excludes fixed costs. The reason for this apparent contradiction is to be found in the limitations of the flexibility requirement, which just requires that the cost function be a local approximation, in some neighborhood of y, but not necessarily at the neighborhood of $y = 0$ which defines the fixed cost. In the sequel we rely on a functional form which is flexible at two points.

Definition 4.4. *A two-points Flexible Functional Form (2FFF) for a cost function provides a second order approximation to an arbitrary twice continuously differentiable cost function C at point where $y > 0$ and at $y = 0^+$.*

We have seen that a production technology with fixed cost, can be represented by *two* different production technologies: one for initiating production $H(x_f) \equiv G(0, x_f)$ (using only fixed inputs), and one for reaching the output level y, and given by $G(x_v, x_f)$. So it becomes quite natural to specify both technologies in a flexible way. Similarly, the cost function is additively separable in two parts: one part u corresponding to the cost at zero output level and one part, v, reflecting the production cost of the output. So if our objective is to provide an approximation of the production technology, both parts should be treated with equal importance, and we suggest here to use a flexible functional form for both the fixed and variable cost functions. Definition 4.4 implies that a 2FFF cost function is the sum of two 1FFF fixed and variable cost functions U and V.

Diewert and Wales (1987, p.45-46) define a one point (1FFF) flexible cost function at the point (w^0, y^0, t^0) as one being able to approximate an arbitrary cost function C^0 locally, where C^0 is continuous and homogeneous of degree one in w. This definition is satisfied if and only if C has "enough free parameters so that the following $1 + (J + 2) + (J + 2)^2$ equations can be satisfied":

$$\begin{aligned} C\left(w^0, y^0, t^0\right) &= C^0\left(w^0, y^0, t^0\right) \\ \nabla C\left(w^0, y^0, t^0\right) &= \nabla C^0\left(w^0, y^0, t^0\right) \\ \nabla^2 C\left(w^0, y^0, t^0\right) &= \nabla^2 C^0\left(w^0, y^0, t^0\right), \end{aligned} \tag{4.20}$$

where the ∇C (respectively $\nabla^2 C$) denotes the first (second) order partial derivatives with respect to all arguments of C. Since the Hessian is symmetric and C is linearly homogeneous in w, this system includes only $J(J + 1)/2 + 2J + 3$ free equations. The requirements (4.20) have to be fulfilled at a single point y^0 which can be chosen to be positive, so the 1FFF definition is compatible with the absence of fixed cost. This

explains why the Translog is flexible although $U \equiv 0$. This drawback of 1FFF explains why we consider 2FFF.

A 2FFF for a cost function has enough free parameters for satisfying the following $1 + (J+1) + (J+1)^2 + 1 + (J+2) + (J+2)^2$ equations:

$$\begin{aligned} U\left(w^0, t^0\right) &= U^0\left(w^0, t^0\right), \\ \nabla U\left(w^0, t^0\right) &= \nabla U^0\left(w^0, t^0\right), \\ \nabla^2 U\left(w^0, t^0\right) &= \nabla^2 U^0\left(w^0, t^0\right), \end{aligned} \tag{4.21}$$

and for $y^0 > 0$,

$$\begin{aligned} V\left(w^0, y^0, t^0\right) &= V^0\left(w^0, y^0, t^0\right), \\ \nabla V\left(w^0, y^0, t^0\right) &= \nabla V^0\left(w^0, y^0, t^0\right), \\ \nabla^2 V\left(w^0, y^0, t^0\right) &= \nabla^2 V^0\left(w^0, y^0, t^0\right). \end{aligned} \tag{4.22}$$

Since U is linearly homogeneous in w, and its Hessian is symmetric, this imposes the following additional restrictions $2 + J + (J+1)\,J/2$ on U:

$$\begin{aligned} w^\top \frac{\partial U}{\partial w}(w,t) &= U(w,t), \qquad w^\top \frac{\partial^2 U}{\partial w \partial t}(w,t) = \frac{\partial U}{\partial t}(w,t), \\ w^\top \frac{\partial^2 U}{\partial w \partial w^\top}(w,t) &= 0, \qquad \nabla^2 U(w,t) = \nabla^2 U(w,t)^\top \end{aligned}$$

It turns out the fixed cost function U has at least $(J+1) + J\,(J+1)\,/2$ free parameters in order to be flexible. Similarly, the variable cost function V must have at least $(J+2) + (J+1)\,(J+2)\,/2$ free parameters. In total, a 2FFF cost function must have at least $1 + 3\,(J+1) + J\,(J+1)$ free parameters. Moreover, in order to identify V as a variable cost function, we impose:

$$V\left(w^0, 0, t^0\right) = 0.$$

Note that (4.21) and (4.22) imply (4.20), but not conversely.

4.6 Econometric treatment of cost heterogeneity

In our most general model, the level of fixed input is not necessarily optimal and has an impact on both the fixed and variable cost:

$$c_r(w, x_f, y, t) = u_r(w, x_f, t) + v_r(w, x_f, y, t),$$

which is somewhat embarrassing as we do not observe the level of x_f, but only total input quantity x. However, our objective is not to estimate firm specific functions v_r and u_r but rather their conditional mean given the value of the observed explanatory variables w, y and t, so we consider:

$$\begin{aligned} V(w,y,t) &\equiv \mathrm{E}\left[v_r(w,x_f,y,t)\,|w,y,t\right], \\ U(w,t) &\equiv \mathrm{E}\left[u_r(w,x_f,t)\,|w,t\right]. \end{aligned}$$

Here integration is over unobserved heterogeneity with respect to the joint distribution of x_f and the individual cost functions v_r and u_r. Using these definitions, we rewrite the model as follow:

$$c_r(w,x_f,y,t) = \gamma^U(w,x_f,t)\,U(w,t) + \gamma^V(w,x_f,y,t)\,V(w,y,t), \tag{4.23}$$

where the functions γ^U and γ^V are defined by:

$$\gamma^U(w,x_f,t) \equiv \frac{u_r(w,x_f,t)}{U(w,t)}, \qquad \gamma^V(w,x_f,y,t) \equiv \frac{v_r(w,x_f,y,t)}{V(w,y,t)},$$

and satisfy $\mathrm{E}\left[\gamma^U|w,t\right] = \mathrm{E}\left[\gamma^V|w,y,t\right] = 1$. Note that the covariance between γ^U and γ^V can *a priori* take any value. However, we derive an important statistical relationship between the fixed and variable cost functions $\gamma^U U$ and $\gamma^V V$.

Proposition 4.5. *Under the assumptions that, (a) individual heterogeneity in the fixed and variable cost functions is independent of x_f; (b) the fixed inputs x_f are positive and are optimally allocated; then:*

(i) *the conditional covariance* $\mathrm{cov}\left(\gamma^U,\gamma^V|w,y,t\right)$ *is non-positive;*

(ii) *the conditional variance matrix* $\mathrm{V}\left[\gamma|w,y,t\right]$ *is singular.*

When the fixed inputs are unobserved we will not be able to estimate functions u_r and v_r, and we cannot test whether $\partial c_r/\partial x_f = 0$ is satisfied or not. However, we will be able to estimate $\mathrm{V}\left[\gamma|w,y,t\right]$ and $\mathrm{cov}\left[\gamma^U,\gamma^V|w,y,t\right]$. If the statistical test leads to rejection of the singularity of $\mathrm{V}\left[\gamma|w,y,t\right]$ or $\mathrm{cov}\left[\gamma^U,\gamma^V|w,y,t\right] \leq 0$, then we can deduce that either the fixed inputs are not optimally allocated (Proposition 4.5), or that the production technology has the specific structure given in (4.16). The level of the fixed cost $\gamma^U U$ and the level of the variable cost $\gamma^V V$ are certainly positively correlated with any dataset: both the fixed and the variable cost increase over time, and firms with a high fixed cost certainly produce more than smaller firms and also have a higher variable cost. Hence the positive correlation between $\gamma^U U$ and $\gamma^V V$. Proposition 4.5, however, states that there is a trade-off – a negative correlation –

between the fixed and the variable cost *for given values* of the explanatory variables (w, y, t). Such a trade-off cannot be directly observed in a dataset, because it pertains to unobserved heterogeneity. With panel data, the issue of interrelated heterogeneity is often discarded, one exception is Gladden and Taber (2009) who considered it in estimating linear wage equations. In contrast to Gladden and Taber (2009), we derive the sign of the covariance from a structural nonlinear model.

Let us now explain our strategy for estimating this covariance along with other statistics of interest. We have to explicitly introduce the parameters in the notations of the cost function and rewrite the observed cost level c_{nt} as follow:

$$c_{it} = \gamma_{it}^{U} U\left(w_{it}, t; \alpha\right) + \gamma_{it}^{V} V\left(w_{it}, y_{it}, t; \beta\right) + e_{it}, \tag{4.24}$$

where $i = 1, \ldots, N$ denotes the sector, $t = 1, \ldots, T$ represents time. The random term e_{it} is i.i.d., satisfies $\mathrm{E}\left[e_{it} | w_{it}, y_{it}, t\right] = 0$ and has constant variance σ_c^2. It is also assumed that e_{it} is uncorrelated with $\gamma_{it} \equiv \left(\gamma_{it}^{U}, \gamma_{it}^{V}\right)^{\top}$ and any right hand side regressors. Equivalently, we can write our empirical model as:

$$c_{it} = U\left(w_{it}, t; \alpha\right) + V\left(w_{it}, y_{it}, t; \beta\right) + \varepsilon_{it}^{c}. \tag{4.25}$$

with the composite error term:

$$\varepsilon_{it}^{c} \equiv \left(\gamma_{it}^{U} - 1\right) U\left(w_{it}, t; \alpha\right) + \left(\gamma_{it}^{V} - 1\right) V\left(w_{it}, y_{it}, t; \beta\right) + e_{it}. \tag{4.26}$$

Note that $\mathrm{E}\left[\varepsilon_{it}^{c} | w_{it}, y_{it}, t\right] = 0$. We also assume that:

$$\mathrm{V}\left[\gamma_{it} | w, y, t\right] \equiv \Sigma = \begin{pmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{pmatrix}, \tag{4.27}$$

and $\mathrm{V}\left[\gamma_{it}\gamma_{js}^{\top} | w, y, t\right] = 0$, for any $i \neq j$ and $t \neq s$. This model is an extension of Swamy's (1970) random coefficient model to our nonlinear setup with individual and time varying random coefficients. The values of γ_{it} can be considered as incidental parameters, because they are not fundamentally interesting (and cannot be identified). Their distribution however is informative. The joint distribution of γ_{it} reflects the way the variable and fixed cost vary together. The covariance between γ_{it}^{U} and γ_{it}^{V} allows to discriminate between the case of optimally and non-optimally allocated fixed input and whether fixed cost has an impact on the marginal cost of production and the markup via (4.17). The parameters of interest are the technology parameters $\theta \equiv \left(\alpha^{\top}, \beta^{\top}\right)^{\top}$ and the variance matrix Σ.

In principle, all estimates of the technology parameters θ and the covariance matrix can be obtained simultaneously by solving (numerically) the likelihood maximization

or the nonlinear least squares problem.[10] However, these objective functions are highly nonlinear in θ, and it turns out that nonlinear numerical algorithms often do not converge to a solution. We avoid these numerical problems, and use a two-stage estimation procedure. First, the technological parameters θ are consistently estimated (without identification of Σ and σ_c^2) by minimizing the sum of squared residuals:

$$\widehat{\theta} = \arg\min_{\alpha,\beta} \sum_{i,t} \left[c_{it} - U\left(w_{it}, t; \alpha\right) - V\left(w_{it}, y_{it}, t; \beta\right)\right]^2.$$

As the random term ε_{it}^c exhibits heteroscedasticity and serial correlation, we rely on the Newey-West (1987) estimator for estimating the variance matrix $\mathrm{V}\left[\widehat{\theta}\right]$.

In the second-stage, two equivalent estimation methods are again available: Maximum Likelihood (ML) and Least Squares (LS). The conditional variance of $\hat{\varepsilon}_{it}^c$ can be expressed as (using (4.26)):

$$\begin{aligned} & \mathrm{E}\left[(\hat{\varepsilon}_{it}^c)^2 \left| w_{it}, y_{it}, t\right.\right] \equiv \triangle_{it}(\sigma_c^2, \Sigma, \widehat{\theta}) \qquad (4.28) \\ = \;& \sigma_c^2 + \sigma_U^2 U^2\left(w_{it}, t; \widehat{\alpha}\right) + \sigma_V^2 V^2\left(w_{it}, y_{it}, t; \widehat{\beta}\right) + 2\sigma_{UV} U\left(w_{it}, t; \widehat{\alpha}\right) V\left(w_{it}, y_{it}, t; \widehat{\beta}\right). \end{aligned}$$

It turns out that the parameters σ_c^2, σ_U^2, σ_V^2 and σ_{UV} of (4.28) can be estimated by an OLS regression of the squared NLS residuals:

$$(\hat{\varepsilon}_{it}^c)^2 = \left[c_{it} - U\left(w_{it}, t; \widehat{\alpha}\right) - V\left(w_{it}, y_{it}, t; \widehat{\beta}\right)\right]^2, \qquad (4.29)$$

on a constant, $\widehat{U}^2$, $\widehat{V}^2$ and $\widehat{U}\widehat{V}$. If we assume that the heterogeneity vector γ_{it} and the error term e_{it} follow some parametric distribution, then the estimated covariance matrix can be obtained by maximizing the likelihood function. Both second-stage estimation methods are asymptotically equivalent, but their estimation outcomes may differ: first, because the ML is more efficient than OLS if the distribution of the random terms is well specified; second, because the covariance matrix Σ is not restricted to be positive-definite in the OLS regression, but this restriction is imposed in most ML estimation algorithms.[11] As this matrix may well be singular (Proposition 4.5), we prefer the OLS

[10] The NLS estimator of $\left(\theta, \sigma_c^2, \Sigma\right)$ could be obtained (in one step) by minimizing the following sum of squared residuals:

$$\sum_{i,t} \left[\varepsilon_{it}^{c2}(\theta) - \sigma^2 - \sigma_U^2 U^2\left(w_{it}, t; \alpha\right) - \sigma_V^2 V^2\left(w_{it}, y_{it}, t; \beta\right) - 2\sigma_{UV} U\left(w_{it}, t; \alpha\right) V\left(w_{it}, y_{it}, t; \beta\right)\right]^2.$$

An alternative estimator of parameters θ, σ_c^2 and Σ is the maximum likelihood estimator. Under the normality assumption of the random term $\varepsilon_{it}^c \sim N\left(0, \triangle_{it}\left(\theta, \sigma_c^2, \Sigma\right)\right)$, we can write

$$\log L(\theta, \sigma_c^2, \Sigma) = -\frac{1}{2} \sum_{i,t} \left\{\log(2\pi) + \log \triangle_{it}\left(\theta, \sigma_c^2, \Sigma\right) + \triangle_{it}\left(\theta, \sigma_c^2, \Sigma\right)^{-1} \left(\varepsilon_{it}^c(\theta)\right)^2\right\}.$$

[11] It can be shown that the first order conditions of ML are identical to the moment conditions of OLS.

approach.

Our estimation approach can be viewed as a sequential two-stage M-estimation, where in the first-stage $\widehat{\theta}$ is obtained by solving a NLS problem and then, given $\widehat{\theta}$, the estimates $\widehat{\sigma}^2, \widehat{\Sigma}$ are obtained by OLS. This second stage estimator is simple and consistent if the first-stage estimator is consistent for θ, see Cameron and Trivedi (2005, Section 6.6). However, the asymptotic distribution of $\hat{\Sigma}$ given the estimation of $\widehat{\theta}$ is difficult to establish. Hence we use the panel bootstrap for deriving the standard deviations of the second-stage estimator.[12]

4.7 Empirical investigation

In this section, we first summarize the empirical models and strategies, we then present briefly the data set and discuss the estimation results.

4.7.1 Empirical models and estimation strategies

For the empirical fixed and variable cost functions U and V, we assume Translog functional forms denoted by U^{TL} and V^{TL}. As seen in Proposition 4.4, the traditional Translog cost function C^{TL} satisfies $C^{TL}(w, 0, t) = 0$ and is not compatible with the occurrence of a fixed cost (in the best case where $\beta_{yy} \leq 0$). It is, however, quite simple to generalize the Translog specification by adding a fixed cost function to the variable Translog cost function (the two-points flexible form):

$$C^{TL}(w, y, t; \alpha, \beta) = U^{TL}(w, t; \alpha) + V^{TL}(w, y, t; \beta),$$

where

$$U^{TL}(w, t; \alpha) = exp\{\alpha_0 + \alpha_w^\top \ln w + \alpha_t t + \frac{1}{2} \ln w^\top A_{ww} \ln w + \ln w^\top A_{wt} t + \frac{1}{2}\alpha_{tt} t^2\}, \quad (4.30)$$

and

$$\begin{aligned} V^{TL}(w, y, t; \beta) &= exp\{\beta_0 + \beta_w^\top \ln w + \beta_y \ln y + \beta_t t + \frac{1}{2} \ln w^\top B_{ww} \ln w \\ &\quad + \ln w^\top B_{wy} \ln y + \ln w^\top B_{wt} t + \frac{1}{2}\beta_{yy} (\ln y)^2 + \beta_{yt} t \ln y + \frac{1}{2}\beta_{tt} t^2\}. \end{aligned}$$

We impose linear homogeneity and symmetry in w using the following $2+J+(J+1)J/2$ parametric restrictions on U^{TL}:

$$\iota^\top \alpha_w = 1, \quad \iota^\top A_{wt} = 0, \quad \iota^\top A_{ww} = 0, \quad A_{ww} = A_{ww}^\top. \quad (4.31)$$

[12] We assume that the errors are i.i.d. over individuals (but not over time). The panel bootstrap performs a classical paired bootstrap that resamples only over i and not over t.

There are $1 + J + (J+1)J/2$ free parameters left in U^{TL}. Similarly, the variable cost function V^{TL} has $3 + 2J + (J+1)J/2$ free parameters which satisfies:

$$\iota^\top \beta_w = 1, \quad \iota^\top B_{wt} = \iota^\top B_{wy} = 0, \quad \iota^\top B_{ww} = 0, \quad B_{ww} = B_{ww}^\top. \tag{4.32}$$

Note that the logarithmic transformation of the total cost function is not useful anymore for linearizing the nonlinear Translog specification (unless $U^{TL} \equiv 0$). For $J = 4$, the fixed cost function has 15 free parameters to which are added the 21 free parameters of the variable cost function.

Given the two-points flexible specification, we estimate the parameters α and β by using NLS based on (4.25) in the first-stage. The second-stage consists in the estimation of the variance matrix Σ and σ_c^2 by using OLS based on (4.28) and (4.29). The classical Translog cost function which includes only the variable cost function V^{TL} (and assumes that $U^{TL} \equiv 0$) is also considered for comparison. We consider further empirical models that include the system estimation by adding the input demand equations (obtained by applying Shephard's lemma to C^{TL}), as well as the model estimated in first-differences. Substantial gains in efficiency can be realized by system estimation, because more observations are available. The first-difference estimation model is more robust against non-stationarity of the series and unobserved individual fixed effects. Henceforth, Model I denotes the single equation model without any fixed cost. Model II is the baseline model where the cost function includes both a fixed and a variable part (the two-points flexible form). More efficient frameworks are Model III (in level) and Model IV (in difference), which include the cost and the input demand functions. We note that the choice of starting values is crucial for reaching the optimum in the case of system NLS.[13]

4.7.2 Data and empirical results

We use the NBER-CES manufacturing industry database for our empirical study.[14] This database records annual information on output y_{it}, output price p_{it}, and the input levels x_{it}, together with input prices indexes w_{it}, for 462 U.S. manufacturing industries (at the six-digit NAICS aggregation level) and covers the period 1958 to 2005. See Chapter 3 for descriptive statistics and details on the computations made for generating the depreciation rate, interest rate, and the user cost of capital. Information is available for four inputs: capital, labor, energy and intermediate materials.

We begin by commenting the first-stage estimation results for models I to IV (Table 4.1). Instead of reporting estimates for all Translog parameters, we only select

[13] For the single equation estimation (Model I and II), the starting values are set arbitrarily to zero. For the system estimation in levels (Model III), the starting values are the estimates obtained from Model II. For the system estimation in first differences (Model IV), the starting values are obtained from the estimation of the cost function in first-differences.

[14] The dataset can be downloaded at: http://www.nber.org/data/nbprod2005.html

some informative estimated coefficients and statistics. An important coefficient is the parameter β_{yy}, which is crucial for Proposition 4.5. Given the estimated Translog coefficients, we compute statistics such as the share of the fixed cost in the total cost U/C, the ratio of the output price to the predicted marginal cost of production $p/\left(\partial C/\partial y\right)$ which measures the markup, and the rate of returns to scale $1/\varepsilon\left(C,y\right)$, where $\varepsilon\left(C,y\right) \equiv \partial \ln C/\partial \ln y$ denotes the elasticity of costs with respect to output.

As mentioned in Section 4.4, neglecting the fixed cost is a source of bias. By comparing the estimation outcomes of Model I and Model II, we note that the results of the two models differ with respect to several key points. First, the parameters of the fixed cost function (α) in Model II are significantly different from zero, which indicates the existence of fixed costs in the production process. Second, the model without a fixed cost (Model I) suggests that the industries exhibit decreasing returns to scale, but the model with a fixed cost (Model II) suggests increasing returns to scale. The bias on the degree of returns to scale is due to the overestimation of the elasticity of cost and neglect of the fixed cost (see Section 4.4). Finally, the overestimation of marginal costs by Model I leads to underestimation of the markup: the median of $p/\left(\partial C/\partial y\right)$ in Model I is about 36% lower than the one predicted by Model II.[15]

Table 4.1 also shows that empirical results obtained from models II to IV exhibit some regularities. First, the estimated coefficient of β_{yy} is significantly negative in all cases, which implies that the limit of the classical Translog variable cost function is zero as y approaches 0. Second, all models predict that the fixed cost represents a considerable share of total cost. The median of estimated shares U/C varies in the range between 51% and 76%. Third, the estimation results also suggest that the industries exhibit increasing returns to scale. The median of the rate of returns to scale, $\varepsilon\left(C,y\right)^{-1}$ ranges between 1.4 and 2.1. Fourth, there is a significant difference between the selling price and the predicted marginal cost of production, the median of estimated markup varies from 1.8 to 2.6. However, we note that the results of Model IV (with data in first-differences) differ quantitatively from those of Model II and Model III (with data expressed in levels).

Now, we focus on the fixed cost share (U/C), in particular on its evolution over time. These series (averaged over all industries) are depicted on Figure 4.5. We note that for all the empirical models, the fixed cost shares are decreasing over time. This may reflect firms' efforts to increase production flexibility. The series generated by model II (where the input demands system is not included in the estimation), exhibit a structural break around 1980. For other models, the decline of fixed cost shares over time is smoother. However, the decrease is less significant in the first-differenced model (Model IV).

When it comes to the second-stage estimation, the estimates of σ_U^2, σ_V^2, σ_{UV} and σ_c^2, are somewhat more divergent across the models. However, we see that the variance

[15] We also reestimate Model I after appending a linear fixed cost term $w^\top \widetilde{\alpha}$ in the specification. The corresponding empirical results are not reported, but lie in between those obtained for Model I and II.

Table 4.1: Summary of estimation results

	Model	I	II	III	IV
U/C	1st q	-	0.19	0.27	0.48
	median	-	0.52	0.51	0.76
	3rd q	-	0.91	0.77	0.94
$p/\left(\partial C/\partial y\right)$	1st q	1.19	1.32	1.37	1.58
	median	1.36	1.87	1.78	2.63
	3rd q	2.47	6.60	2.74	5.73
$1/\varepsilon\left(C,y\right)$	1st q	0.69	1.01	1.01	1.10
	median	0.89	1.40	1.37	2.07
	3rd q	0.98	6.04	2.71	7.07
β_{yy}	coeff	-0.05	-0.14	-0.15	-0.32
	t-value	-2.05	-4.21	-3.11	-3.85
σ_U^2	coeff	-	0.54	30.83	1.09
	t-value	-	1.67	1.29	0.15
σ_V^2	coeff	-	0.02	0.27	0.12
	t-value	-	1.95	1.96	1.27
σ_{UV}	coeff	-	-0.32	-13.04	-1.19
	t-value	-	-0.91	-1.43	-0.11
σ_c^2	coeff	3.7e+6	1.9e+6	1.3e+6	2.1e+6
	t-value	-	2.09	0.24	0.66

Notes: Rows 2 to 11 report the estimated parameter values and the corresponding t-statistic for the hypothesis that the parameter is equal to zero. Rows 12 to 20 report the median value of the corresponding statistic over all observations as well as the 1st and 3rd quartiles.

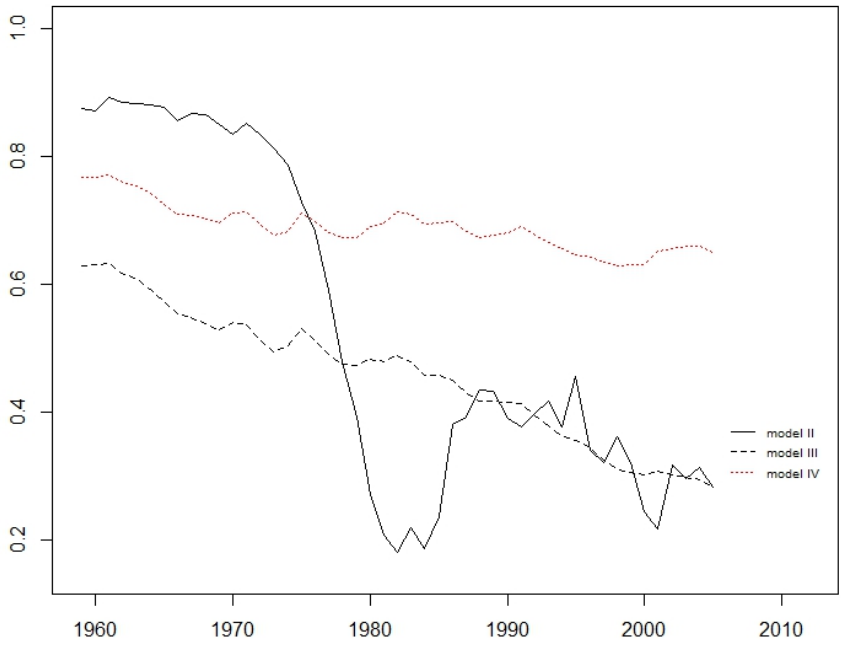


Figure 4.5: Fixed costs shares over time

of the fixed cost heterogeneity γ^U is always larger than the variance of the variable cost heterogeneity γ^V. The covariance between heterogeneities is found to be negative and the covariance matrix Σ is close to singular for all models, which is conform to what we expect from Proposition 4.6. The second-stage estimation results, however, are not precisely estimated and are not statistically significant. This result may be due to our overly restrictive assumption of random heterogeneity in the fixed and variable cost function specification (4.23). Economically, this heterogeneity may well be correlated with further explanatory variables which are individual specific (like for instance the level of production, the type of industry, etc.). So we conduct further analysis in the next subsection.

4.7.3 Estimation with industry specific dummies

Although Models II to IV with random heterogeneity yield some interesting results on the scope of fixed cost and returns to scale, the interaction between fixed and variable costs was not precisely estimated. This may due to the fact that heterogeneity is not purely random but correlated with sectoral characteristics as the level of production or the level of fixed and variable cost. We pursue the investigation a step further and introduce individual-specific dummies into Model IV. The most flexible specification

replaces γ_{it}^U and γ_{it}^V in regression (4.24) by $2N$ individual-specific parameters. In order to limit over-parameterization, we introduce instead dummies for more broadly defined groups of industries. There are different ways to define these groups, for instance, in the spirit of Mundlak's (1978) correlated random coefficient model, individuals can be grouped w.r.t. the average value of their covariates. For the industry database, however, a more natural clustering criterion, is to group the 462 manufacturing sectors available at the six-digit NAICS level into 20 three-digit NAICS sectors. See Table 4.2 for a list of the 3-digit industries.[16]

Formally, we introduce the multiplicative dummy variables γ_j^U and γ_j^V for $j = 1, ..., 20$ in place of the random parameters of (4.24) which becomes:

$$c_{it} = \gamma_j^U U^{TL}(w_{it}, t; \alpha) + \gamma_j^V V^{TL}(w_{it}, y_{it}, t; \beta) + e_{it}. \tag{4.33}$$

Since the Translog cost function also includes the terms α_0 and β_0, all the parameters cannot be identified separately, unless we consider two additional restrictions. Since by construction, we have $\mathrm{E}\left[\gamma^U|w,t\right] = \mathrm{E}\left[\gamma^V|w,y,t\right] = 1$, it is natural to impose the normalization conditions:

$$\frac{1}{20}\sum_{j=1}^{20}\gamma_j^U = \frac{1}{20}\sum_{j=1}^{20}\gamma_j^V = 1,$$

which allow to identify all parameters. In this case, the estimated parameters γ_j^U and γ_j^V represent the industry-specific deviation in percentage from the average. For instance, if the estimated value of γ_j^U is significantly above one and the estimated value of γ_j^V is significantly below one, this indicates that the industry group j incurs more fixed and less variable costs than average. In this framework, the interaction between the fixed and variable components of the cost function is characterized by the variation of γ_j^U and γ_j^V over industry groups. We examine the empirical correlation between γ_j^U and γ_j^V along with group-specific shares of fixed cost, degree of returns to scale, markups and rate of technical change.

We estimate the parameters of the extended Model IV and report the estimation results in Table 4.2. Column 3 and 4 of Table 4.2 report the estimated coefficients of γ_j^U and γ_j^V. Our estimation results indicate, for instance, that compared to the average, the industry group NAICS 311 (food) operates with 24% less fixed cost and 4% less variable cost than the average. We also note that industries with lower than average fixed cost generally have higher than average variable cost and conversely. Contrary to the above random effect models, the parameters reflecting cost heterogeneity are now statistically significant.

Columns 5 to 10 report the median (for each group) of the fixed cost share U/C,

[16] At the three-digit NAICS level, there are actually 21 manufacturing industry groups. We merge the smallest (in terms of the number of subsectors) NAICS 324 industry group (petroleum and coal products manufacturing) with NAICS 325 industry group (chemical manufacturing).

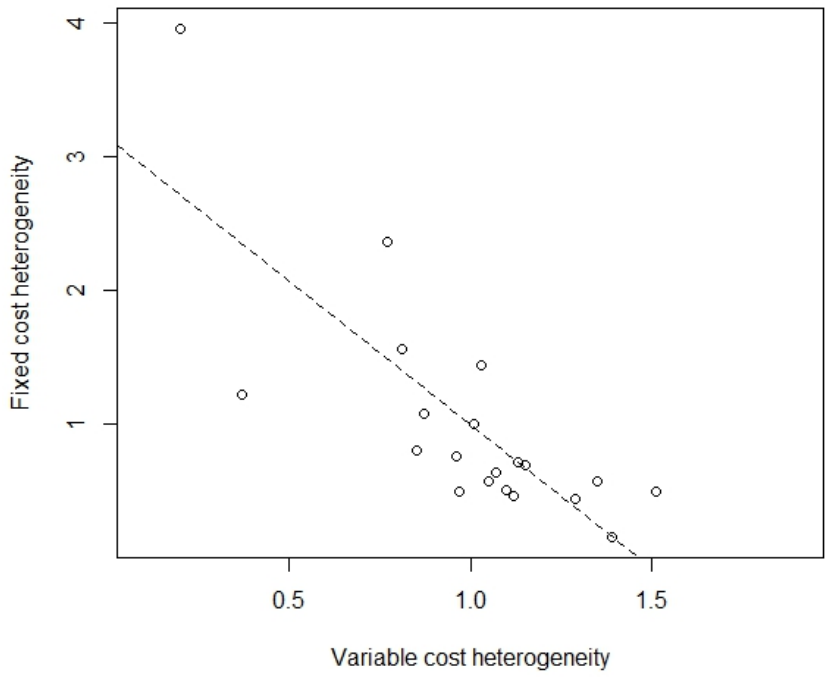


Figure 4.6: Scatterplot of $\widehat{\gamma}_j^U$ and $\widehat{\gamma}_j^V$

the markup $p/\left(\partial C/\partial y\right)$, returns to scale $1/\varepsilon\left(C,y\right)$, and technical change measured as $\partial \ln C/\partial t$, $\partial \ln U/\partial t$ and $\partial \ln V/\partial t$. For the NAICS 311 industry group, the estimates indicate that the fixed cost represents 25% of the total production cost, with almost constant returns to scale and a markup of 68%. In average over all industries, the results confirm former findings with strong evidence for fixed cost, increasing returns to scale and markup pricing. We also find evidence for the conjecture brought forward in Section 4.4: industries with higher fixed cost also exhibit higher markups and returns to scale.

Regarding the degree of technical change, our results on $\partial \ln V/\partial t$ show that the variable cost is on average decreasing by 0.9% over time with little variance over industries. In contrast, the fixed cost increases with time i.e., $\partial \ln U/\partial t = 0.04$. Altogether, our results are in line with those obtained by Diewert and Fox (2008) who found modest empirical evidence for technical change in U.S. manufacturing. Our interpretation is that the deterministic trend only partially captures technical progress, and that one important part of technical change is stochastic and embodied in the unobserved fixed inputs (the x_f). These fixed inputs contribute to increase the fixed cost and decrease the variable cost and, as a consequence of our approach, this random component of technical change is captured by the negative correlation between γ_j^U and γ_j^V.

Table 4.2: Summary of estimation results with industry dummies

NAICS	industry groups	γ_j^U	γ_j^V	$\gamma_j^U U/C$	$p/(\partial C/\partial y)$	$1/\varepsilon(C,y)$	$\partial \ln C/\partial t$	$\partial \ln U/\partial t$	$\partial \ln V/\partial t$
311	Food	0.76 (7.36)	0.96 (26.40)	0.25	1.68	1.05	0.004	0.039	-0.009
312	Beve.&Toba.	1.22 (6.03)	0.37 (6.22)	0.62	4.42	2.07	0.020	0.037	-0.009
313	Textile	0.58 (5.02)	1.35 (18.15)	0.34	1.57	1.11	0.008	0.041	-0.009
314	Textile Prod.	0.51 (7.17)	1.10 (17.52)	0.41	2.14	1.19	0.015	0.042	-0.009
315	Apparel	0.50 (6.55)	1.51 (18.64)	0.28	1.57	1.01	0.008	0.044	-0.009
316	Leather	0.16 (3.76)	1.39 (11.83)	0.27	2.18	0.97	0.005	0.042	-0.010
321	Wood	1.00 (10.22)	1.01 (24.23)	0.41	1.74	1.29	0.009	0.038	-0.009
322	Paper	1.56 (10.03)	0.81 (16.49)	0.61	2.15	1.93	0.019	0.036	-0.010
323	Printing	0.70 (3.16)	1.15 (8.52)	0.31	1.50	1.13	0.005	0.038	-0.009
324-5	Petr.&Chem.	0.46 (2.91)	1.12 (24.15)	0.17	1.35	0.96	-0.001	0.035	-0.009
326	Plastic	1.44 (8.86)	1.03 (17.43)	0.52	1.74	1.62	0.012	0.037	-0.009
327	Mineral Prod.	0.64 (9.94)	1.07 (26.75)	0.43	1.83	1.31	0.011	0.036	-0.010
331	Primary Metal	1.08 (6.91)	0.87 (11.44)	0.47	1.85	1.46	0.011	0.036	-0.010
332	Fabricated Metal	0.72 (8.73)	1.13 (28.03)	0.32	1.47	1.13	0.006	0.037	-0.009
333	Machinery	0.81 (14.34)	0.85 (26.75)	0.42	1.88	1.33	0.010	0.036	-0.010
334	Computer	3.96 (9.62)	0.20 (4.10)	0.94	8.81	13.47	0.037	0.042	-0.009
335	Elec.Equipement	0.58 (7.36)	1.05 (32.20)	0.30	1.70	1.10	0.004	0.038	-0.009
336	Transportation	2.37 (10.41)	0.77 (26.46)	0.49	1.84	1.61	0.011	0.035	-0.009
337	Furniture	0.44 (5.06)	1.29 (22.58)	0.28	1.50	1.04	0.004	0.040	-0.009
338	Miscellaneous	0.50 (7.29)	0.97 (15.30)	0.45	2.16	1.32	0.014	0.040	-0.009
	Average	1.00	1.00	0.42	2.25	1.91	0.011	0.039	-0.009

Note: t-values are reported in parenthesis for $H_0 : \gamma_j^U = 0$ and for $H_0 : \gamma_j^V = 0$.

Table 4.3: Correlation matrix

	γ_j^U	γ_j^V	$\gamma_j^U \frac{U}{C}$	$\frac{p}{\partial C/\partial y}$	$\frac{1}{\varepsilon(C,y)}$	$\frac{\partial \ln C}{\partial t}$	$\frac{\partial \ln U}{\partial t}$	$\frac{\partial \ln V}{\partial t}$	$\overline{y}_j$	cr_j
γ_j^V	-0.79									
$\gamma_j^U \frac{U}{C}$	0.86	-0.84								
$\frac{p}{\partial C/\partial y}$	0.80	-0.77	0.83							
$\frac{1}{\varepsilon(C,y)}$	0.86	-0.67	0.79	0.95						
$\frac{\partial \ln C}{\partial t}$	0.81	-0.78	0.97	0.88	0.83					
$\frac{\partial \ln U}{\partial t}$	-0.07	0.31	0.01	0.28	0.26	0.20				
$\frac{\partial \ln V}{\partial t}$	0.22	0.02	-0.03	0.26	0.29	0.09	0.56			
$\overline{y}_j$	0.58	-0.58	0.29	0.27	0.28	0.18	-0.55	0.22		
cr_j	0.23	-0.39	0.20	0.23	0.07	0.20	-0.16	0.14	0.58	
H_j	0.24	-0.45	0.25	0.21	0.01	0.24	-0.20	0.17	0.61	0.92

Table 4.3 reports the empirical correlations between different estimated statistics. The main result is that the correlation between $\widehat{\gamma}_j^U$ and $\widehat{\gamma}_j^V$ is negative, and quite strong (-0.79). The scatterplot of $\widehat{\gamma}_j^U$ and $\widehat{\gamma}_j^V$ is depicted on Figure 4.6. These results are in line with Proposition 4.5. The extension of Model IV to include industry-specific fixed and variable cost heterogeneity now allows us to find more precise empirical results than those obtained with random heterogeneity. The separable structure of Proposition 4.3, (4.15), which implies no interaction between fixed and variable cost, is statistically rejected: technology $G(x_v, x_f)$ fits the data better than $F(x_v + K(x_f))$ for any function K.

We also find that the fixed-cost heterogeneity is positively correlated with most of the statistics especially with the markup and the rate of returns. This coincides with our discussion of Section 4.4 on the dangers of neglecting fixed cost. Not surprisingly, the correlations involving γ_j^V have the opposite sign to those involving γ_j^U. The strong positive correlation between $\gamma_j^U U/C$ and $p/(\partial C/\partial y)$ seems to be contrary to the prediction made by the theory of contestable markets (Baumol et al., 1988). However, it can be explained in the light of our framework: a higher fixed cost reduces the variable cost (at given level of production), a relationship which is reflected by the negative correlation between γ_j^U and γ_j^V. This negative correlation is in turn inherited by $\gamma_j^U U/C$ and $\partial C/\partial y$.

These results help to understand why specifications neglecting the fixed cost (or including an inflexible parameterization of the fixed cost) are likely to overestimate the marginal cost of production and underestimate the markup and the rate of returns to scale. The omission of the fixed cost leads to attribute neglected variations in fixed costs (which according to Table 4.3 are positively correlated with output) to the variable cost function which is increasing in y. Like in the case of an omitted variable bias, the variable cost function (and especially its partial derivative w.r.t. y) will catch up the

part of the fixed cost function which is correlated with production and so, it will be biased upwards. The positive correlation $corr\left(\gamma_j^U;\overline{y}_j\right)=0.58$ explains the gap between the results obtained with the standard and extended Translog specifications (see Table 4.1). The neglected fixed cost with Model I is directly responsible for low rate of returns to scale and moderate markups obtained with this specification.

Regarding technical change, we find that $\partial \ln C/\partial t$ is positive and highly correlated with γ_j^U, γ_j^V, $\gamma_j^U U/C$ and $p/\left(\partial C/\partial y\right)$, which means that fixed cost and market power preclude productivity growth (as in Arrow, 1962). Surprisingly, neither $\partial \ln U/\partial t$ nor $\partial \ln V/\partial t$ are strongly correlated with market power. This paradox is solved if we go back to the definition of technical change, in which the share of fixed cost plays an important role:

$$\frac{\partial \ln C_j}{\partial t}=\frac{\partial \ln U}{\partial t}\frac{\gamma_j^U U}{C_j}+\frac{\partial \ln V}{\partial t}\left(1-\frac{\gamma_j^U U}{C}\right),$$

and introduces correlation between $\partial \ln C_j/\partial t$ and γ_j^U and $\gamma_j^U U/C_j$. We also investigate the link between the fixed cost, the size and the concentration of industries. Table 4.3 also reports correlations between the fixed cost and the average output level (over time and subsectors within industry j), the concentration ratio for the 20 largest firms cr_j, and the Hirschman-Herfindahl index H_j.[17] We find a positive correlation between the fixed cost share and the industrial concentration. These results suggest that industries with a higher fixed cost and a lower variable cost, produce more in average, and are more concentrated.

4.8 Conclusion

This chapter investigates technologies in which fixed inputs can be imperfectly substituted to variable inputs, and we propose extended production and cost functions compatible with the occurrence of a fixed cost. Many available flexible specifications, like the Translog cost function, restrict the fixed cost to be equal to zero. Our extended specification of the Translog is compatible with arbitrary levels of fixed cost, and allows for interactions between the fixed and the variable cost. Our empirical findings highlight the importance of fixed cost which represent about 20% to 60% of total cost in the manufacturing industries and tend to decline to decline over time. Our estimates also supports our extended framework which explains why industries with higher fixed cost, in average have lower variable cost, higher returns to scale and markups. Conformably to our theoretical prediction, we also find that the classical Translog cost function underestimates the rate of return to scale and the markup.

A natural extension of our framework would be to examine explicitly strategic interactions between firms in their joint decision on product price and production capacity

[17] The concentration data for 2002 are obtained from the U.S. Census Bureau.

(fixed cost). This would potentially allow to revisit the link between fixed cost and barriers to entry.

4.9 Appendix

Proof of Proposition 4.1.

From the definition of X_F and $X_F \neq \varnothing$ it directly follows that:

$$c_r(w, x_f, 0) = \min_{x_v \geq 0} \left\{ w^\top x_v + w^\top x_f : F(x_v + x_f) \geq 0 \right\} = w^\top x_f \geq 0,$$

and so $v_r(w, x_f, 0) = c_r(w, x_f, 0) - w^\top x_f = 0$.

(i) The variable inputs must satisfy the non-negativity constraints $x_v \geq 0$. If these constraints are not binding at the optimum, we can write:

$$c_r(w, x_f, y) = \min_{x_v > 0} \left\{ w^\top x_v + w^\top x_f : F(x_v + x_f) \geq y \right\} = v_r(w, x_f, y) + w^\top x_f,$$

where $v_r(w, x_f, y) \equiv \min_{x > x_f} \left\{ w^\top x : F(x) \geq y \right\} - w^\top x_f > 0$. Then $c_r(w, x_f, y) = C(w, y)$ and by Shephard's lemma $x_v^*(w, x_f, y) = X_v^*(w, y)$.

(ii) If some constraints $x_{v,j} \geq 0$ are binding at the optimum, the total input x can be rewritten as:

$$x = x_v + x_f = \begin{pmatrix} \widetilde{x} \\ \overline{x} \end{pmatrix},$$

with $\widetilde{x}_i = x_{v,i} + x_{f,i}$ for $x_{v,i} > 0$ and $\overline{x}_j = x_{f,j}$ for $x_{v,j} = 0$. Vector w is partitioned accordingly as $w = \left(\widetilde{w}^\top, \overline{w}^\top \right)^\top$. Then

$$\begin{aligned} c_r(w, x_f, y) &= \min_{x_v \geq 0} \left\{ w^\top x_v + w^\top x_f : F(x_v + x_f) \geq y \right\} \\ &= \min_{\widetilde{x} > 0} \left\{ \widetilde{w}^\top \widetilde{x} + \overline{w}^\top \overline{x} : F(\widetilde{x}, \overline{x}) \geq y \right\} \\ &= \min_{\widetilde{x} > 0} \left\{ \widetilde{w}^\top \widetilde{x} : F(\widetilde{x}, \overline{x}) \geq y \right\} + \overline{w}^\top \overline{x} = V_r(\widetilde{w}, \overline{x}, y) + \overline{w}^\top \overline{x}. \end{aligned}$$

Proof of Proposition 4.2.

(i) If $x_f \in X_G$ then $x_v = 0$ is admissible and so:

$$v_r(w, x_f, 0) = \min_{x_v \geq 0} \left\{ w^\top x_v : G(x_v, x_f) \geq 0 \right\} = 0.$$

The assumption that G is single valued and increasing implies that $G(x_v, x_f) > 0$ for and $x_v > 0$ and $x_f \in X_G$. Then $v_r(w, x_f, y) = w^\top x_v^*(w, x_f, y) > 0$ for $y > 0$ because $w > 0$ at least one element of $x_v^*(w, x_f, y)$ is strictly positive.

(ii) For $y' > y$, and G increasing in x_f, it implies that $\{x_v : G(x_v, x_f) \geq y'\} \subset \{x_v : G(x_v, x_f) \geq y\}$ and as a consequence:

$$v_r(w, x_f, y') = \min_{x_v \geq 0} \left\{ w^\top x_v : G(x_v, x_f) \geq y' \right\} > v_r(w, x_f, y).$$

(iii) Similarly, $x'_f > x_f$ and G increasing in (x_v, x_f), implies that $\{x_v : G(x_v, x_f) \geq y\} \subset \left\{ x_v : G\left(x_v, x'_f\right) \geq y \right\}$ and as a consequence:

$$v_r\left(w, x'_f, y\right) = \min_{x_v \geq 0} \left\{ w^\top x_v : G\left(x_v, x'_f\right) \geq y \right\} < v_r(w, x_f, y).$$

Proof of Proposition 4.3.

Part (i), Necessity. For an exogenous level of $x_f \in X_G$, we have:

$$\begin{aligned} v_r(w, x_f, y) &= \min_{x_v \geq 0} \left\{ w^\top x_v : y = F(x_v + K(x_f)) \right\} \\ &= \min_{x_v \geq 0} \left\{ w^\top x_v + w^\top K(x_f) : y = F(x_v + K(x_f)) \right\} - w^\top K(x_f) \\ &= \min_{X \geq K(x_f)} \left\{ w^\top X : y = F(X) \right\} - w^\top K(x_f) \\ &= v_y(w, y) - w^\top K(x_f). \end{aligned}$$

The last line follows from our assumption that $x_v^*(w, y) > 0$ at the optimum. Defining $v(w, y) \equiv v_y(w, y) - v_y(w, 0^+)$ ensures that $v(w, 0^+) = 0$. Defining $u_r(w, x_f) \equiv v_y(w, 0^+) - w^\top K(x_f) + w^\top x_f$ ensures that $c_r(w, x_f, y) = u_r(w, x_f) + v(w, y)$.

Conversely, we can recover the convex hull of all inputs producing y, for a given level of x_f, by solving:

$$\min_w \left\{ w^\top x_v - v_y(w, y) + w^\top K(x_f) \right\}.$$

The corresponding J first order conditions for an inner solution are given by:

$$x_v + K(x_f) - \frac{\partial v_y}{\partial w}(w, y) = 0,$$

which can be solved with respect to w/w_J and y to obtain:

$$y = F(x_v + K(x_f)).$$

If G is quasi-concave in x_v, this convex hull corresponds to the isoquants of G.

Part (ii). Necessity. With (4.15), the first order conditions for an inner solution in

x_f to the cost minimization problem are given by:

$$\frac{\partial u_r}{\partial x_f}(w, x_f) = w,$$

and do not depend on y and so the solutions $x_f^*(w)$. With (4.16), the first order conditions for an inner solution in x_v are:

$$\begin{aligned} w &= \lambda \frac{\partial F}{\partial x_v}(x_v + K(x_f)) \\ y &= F(x_v + K(x_f)), \end{aligned}$$

where λ denotes the Lagrange multiplier. The solution in x_v to this system takes the form $x_v^*(w, x_f, y) = X^*(w, y) - K(x_f)$ and so the restricted cost function (4.15), with $v_y(w, y) \equiv w^\top X^*(w, y)$ and $u_r(w, x_f) = w^\top x_f - w^\top K(x_f)$. Then x_f^* is independent of y.

Sufficiency. If x_f^* depends only upon w, then the first order conditions for an inner solution, given by:

$$\frac{\partial u_r}{\partial x_f}(w, x_f) + \frac{\partial v_r}{\partial x_f}(w, x_f, y) = 0$$

imply that:

$$\frac{\partial^2 v_r}{\partial x_f \partial y}(w, x_f, y) = 0$$

and so $c_r(w, x_f, y) = u_r(w, x_f) + v(w, y)$.

Proof of Proposition 4.4.

We rewrite C^{TL} as:

$$C^{TL}(w, y, t) = b(w, t)\, y^{\beta_y + \ln w^\top B_{wy} + \frac{1}{2}\beta_{yy} \ln y + \beta_{yt} t},$$

with

$$b(w, t) \equiv \exp\left(\beta_0 + \beta_w^\top \ln w + \beta_t t + \frac{1}{2} \ln w^\top B_{ww} \ln w + \ln w^\top B_{wt} t + \frac{1}{2}\beta_{tt} t^2\right) > 0.$$

If $\beta_{yy} \leq 0$, then:

$$\lim_{y \to 0^+} C^{TL}(w, y, t) = 0, \tag{4.34}$$

whereas if $\beta_{yy} > 0$,

$$\lim_{y \to 0^+} C^{TL}(w, y, t) = +\infty.$$

The cost function is nondecreasing in $y > 0$ iff:

$$\frac{\partial C^{TL}}{\partial y}(w, y, t) = \left(\beta_y + \ln w^\top B_{wy} + \beta_{yy} \ln y + \beta_{yt} t\right) \frac{C^{TL}(w, y, t)}{y} \geq 0.$$

If $\beta_{yy} > 0$, then:

$$\lim_{y \to 0^+} \frac{\partial C^{TL}}{\partial y}(w, y, t) < 0,$$

and $\partial C^{TL}/\partial y$ becomes positive only for y sufficiently large.

Proof of Proposition 4.5.

There are two types of unobserved heterogeneities here: one due to unobserved x_f and one due to heterogeneous functional forms for u_r and v_r over individuals. For simplicity we use the subscript r for denoting this heterogeneity. Let $f_{u|x}$ denote the conditional density function of $u_r(w, x_f, t)|x_f$. Under Assumption (a) we can write $f_{u|x} = f_u$ where f_u denotes the marginal density of u_r. Let us define the average fixed and variable cost functions (over all firms in our sample) as:

$$\begin{aligned} \overline{u}(w, x_f, t) &\equiv \int u_r(w, x_f, t) f_u(r) dr \\ \overline{v}(w, x_f, y, t) &\equiv \int v_r(w, x_f, y, t) f_v(r) dr. \end{aligned}$$

These functions still depend on the unobserved heterogeneity in x_f, but individual heterogeneity in the cost functions u_r and v_r has been integrated out. Let us also consider:

$$\overline{\gamma}^U(w, x_f, t) \equiv \frac{\overline{u}(w, x_f, t)}{U(w, t)}, \qquad \overline{\gamma}^V(w, x_f, y, t) \equiv \frac{\overline{v}(w, x_f, y, t)}{V(w, y, t)},$$

and (we skip the arguments for simplicity)

$$\overline{c} = \overline{\gamma}^U U + \overline{\gamma}^V V.$$

Using the optimality condition $\partial c_r/\partial x_f = 0$, and Assumption (a), it follows that $\partial \overline{c}/\partial x_f = 0$. So, conditionally on observations (w, y, t), we write:

$$\begin{aligned} \text{cov}\left[\overline{\gamma}^U, \overline{\gamma}^V\right] &= \text{cov}\left[\left(\overline{c} - \overline{\gamma}^V V\right)/U, \overline{\gamma}^V\right] = \text{cov}\left[-\overline{\gamma}^V V/U, \overline{\gamma}^V\right] = -\frac{V}{U}\text{V}\left[\overline{\gamma}^V\right] \leq 0 \\ \text{V}\left[\overline{\gamma}^U\right] &= \text{V}\left[\left(\overline{c} - \overline{\gamma}^V V\right)/U\right] = \frac{V^2}{U^2}\text{V}\left[\overline{\gamma}^V\right]. \end{aligned}$$

(i) Under Assumption (a) we can write:

$$\begin{aligned} \text{cov}\left(\overline{\gamma}^U, \overline{\gamma}^V\right) &= \int \left(\overline{\gamma}^U - 1\right)\left(\overline{\gamma}^V - 1\right) f_x dx_f \\ &= \int \left(\int_{\mathcal{R}} \gamma^U f_{uv}(r) dr - 1\right)\left(\int_{\mathcal{R}} \gamma^V f_{uv}(r) dr - 1\right) f_x dx_f \\ &= \int \int_{\mathcal{R}} \left(\gamma^U - 1\right)\left(\gamma^V - 1\right) f_{uv}(r) f_x(x_f) dr dx_f \end{aligned}$$

$$
\begin{aligned}
&= \int\int_{\mathcal{R}} \left(\gamma^U - 1\right)\left(\gamma^V - 1\right) f_{uv|x}\left(r|x_f\right) f_x\left(x_f\right) dr dx_f \\
&= \operatorname{cov}\left(\gamma^U, \gamma^V\right),
\end{aligned}
$$

where the fourth equality follows from the fact that under Assumption (a) we have the independence of individual heterogeneity with respect to the level of fixed inputs: $f_{uv|x}\left(r|x_f\right) = f_{uv}\left(r\right)$. Putting things together, we have $\operatorname{cov}(\overline{\gamma}^U, \overline{\gamma}^V) = \operatorname{cov}\left(\gamma^U, \gamma^V\right) \leq 0$.

(ii) Similarly, the variance matrices satisfy $\mathrm{V}\left[\overline{\gamma}\right] = \mathrm{V}\left[\gamma\right]$ and so:

$$
\mathrm{V}\left[\gamma\right] = \begin{bmatrix} \frac{V^2}{U^2}\mathrm{V}\left[\overline{\gamma}^V\right] & -\frac{V}{U}\mathrm{V}\left[\overline{\gamma}^V\right] \\ -\frac{V}{U}\mathrm{V}\left[\overline{\gamma}^V\right] & \mathrm{V}\left[\overline{\gamma}^V\right] \end{bmatrix},
$$

whose determinant is zero.

Chapter 5

Productivity, Fixed Cost and Export[1]

5.1 Introduction

During the past decade, research in the field of international trade has moved from a countries and sector perspective to a firms perspective. The usage of firm-level data has been democratized, which in turn offers new insights on trade behavior. Empirical evidence from firm-level data shows that more productive, larger and more capital intensive firms have a higher probability to become exporters, see for example Bernard and Jensen (1999). In parallel, a series of pioneer works by Baldwin (1988, 1989), Krugman (1989) and Roberts and Tybout (1997) introduce a sunk cost for entering into the export market. In response to these findings, the seminal paper of Melitz (2003) provides a highly tractable theoretical framework for modeling firms' export decisions, in which heterogeneous firms face sunk costs of entry and uncertainty concerning their productivity. However, the drawbacks of his model are that all firms face the same entry cost and heterogeneity only appears in Total Factor Productivity (TFP). From an empirical perspective, the econometric models treat the sunk cost of entry as a common parameter across firms and focus only on testing the existence of entry costs, see Roberts and Tybout (1997), Campa (2004) and Bernard and Jensen (2004). Thus, the heterogeneity of entry costs is largely ignored in this literature.

In this chapter, we propose a Melitz-type model with a heterogeneous entry cost for export markets. This heterogeneity is introduced into the cost structure through productivity, where the entry cost is modeled as a function of productivity. The underlying assumption is that the entry into export markets is less costly for more productive firms. The effect of productivity on entry costs is characterized by a constant elasticity (*the productivity elasticity of entry costs*) in our model. The implication of this assumption

[1]This chapter has been circulated under the title "Self-selection into export market: Does productivity affect entry costs?", Chen and Olland (2013).

is that the minimum entry requirement at equilibrium also depends on the productivity elasticity of entry costs.[2] We find that in a relatively selective export market, a higher degree of dependence between productivity and entry costs yields a lower entry requirement. Conversely, in a relatively open market, the productivity elasticity of entry costs plays the opposite role: a higher dependence between productivity and entry costs increases the entry requirement. In order to test our working assumption that entry costs are affected by productivity, we develop an empirical strategy based on a treatment evaluation model for measuring entry costs, and for evaluating the relationship between entry costs and neutral as well as non-neutral productivity. Our study sheds light on determinants of the entry barriers in the international market and how entry barriers could be reduced from the firms' perspective.

The distinction between various types of fixed costs (per-period or sunk), firm-level differences in fixed costs and their impact on the market structure and trade behavior are fundamental issues but are largely neglected in both theoretical and empirical models. For example, the basic Melitz model suggests that firms may have different marginal cost structures, but share the same operating fixed cost and the same sunk cost of entry, no matter how different they are in terms of productivity. In addition, neither operating nor entry fixed costs are affected by productivity (because the additive separability between productivity and fixed costs is imposed in the production technology). Numerous theoretical reasons and empirical evidence show that fixed cost structures may differ at the firm level, see Chapter 4. For instance, the sunk cost associated with product adaptation and promotion (for the international market) may depend on firm specific characteristics, such as innovation capacity and management skills. Considering that these characteristics are the key elements of productivity, we propose a trade model where firm's productivity partly determines the cost of entry. In such a way, we add firm-level heterogeneity into the fixed cost structure. We point out that the productivity elasticity of entry costs is a crucial parameter for entry conditions at equilibrium. It adds (in comparison to the traditional Melitz model) an indirect channel through which the level of productivity determines the self-selection. Our empirical investigations are focused on French manufacturing firms for the last seven years, where we find significant costs of entry, as well as a relationship between productivity and these costs.

The theoretical models proposed by Krugman (1989), Dixit (1989a,b), Melitz (2003) and by Bernard et al. (2003) gave birth to a large number of empirical studies testing entry cost effects. Roberts and Tybout (1997) develop an empirical model of exporting decision with sunk costs. This dynamic discrete-choice framework quantifies the impacts of sunk cost hysteresis, by directly analyzing the firm's entry and exit patterns (the so-called, *direct approach*).[3] Based on Colombian manufacturing plants data over

[2]The minimum entry requirement in the Melity-type trade model is defined in terms of productivity, and only the most productive firms can enter into the international market.

[3]Before the Robert and Tybout's (1997) model, the empirical investigations have rather focused on

the period 1981-1989, Roberts and Tybout (1997) provide strong evidence for the existence of sunk entry costs. Campa (2004) validates the sunk cost hysteresis assumption for Spanish manufacturing plants. In addition, he also finds that sunk costs of entry into the foreign market are much larger than costs of exit. Coinciding with a growing body of trade literature that emphasizes the role of heterogeneity, Bernard and Jensen (2004) extend Robert and Tybout's dynamic export decision model by adding individual effects. They find that sunk costs of entry are significant for U.S. manufacturing plants over the period 1984-1992, and that plant characteristics reflecting past performance increase the probability of entry. However, a surprising result documented in Bernard and Jensen (2004) is that Total Factor Productivity has no significant impact on trade decisions. This result may be due to their empirical model specification. Similar to Roberts and Tybout (1997) and Campa (2004), the dependent variable in that study is the export participation (a binary variable), and the volume of export is not considered. This approach can only capture one aspect of firms' trade behavior. A further contribution to the literature of sunk entry costs is made by Das et al. (2007), whose empirical results provide a series of policy recommendations for export-oriented reforms. Based on a dynamic structural model, their empirical investigation incorporates both aspects of trade behavior: the participation and exports volume. Their simulations point out that policies targeting export revenues are much more effective than entry cost subsidies. Compared to previous empirical papers on the sunk entry cost, our empirical investigation differs in two aspects: first, our objective is not only to test the existence of sunk costs but also to test the relationship between the sunk cost of entry and productivity. Second, we adopt a different and more flexible empirical methodology using the treatment evaluation model for estimating sunk entry costs and for revealing how firm-specific characteristics influence the cost of entry.

The chapter is organized as follows: in the next section, we extend the Melitz model and discuss the implications of heterogeneous sunk entry costs on the equilibrium. Section 5.3 describes our empirical methodology for measuring the effects of firms' characteristics on the entry cost. Data and estimation results are presented in Section 5.4. Section 5.5 concludes.

5.2 Theoretical model with heterogeneous entry costs

In this section, we present a Melitz trade model with heterogeneous entry costs and their implications on the open economy equilibrium. We consider a simple two-countries framework where firms use labor as the unique factor to produce differentiated products in a monopolistic competition market (see for example, Melitz, 2003, Helpman et al., 2004 or Chaney, 2008). Since productivity is drawn exogenously and does not

asymmetries in the responses of trade flows to exchange rate variations (the so-called, indirect approach).

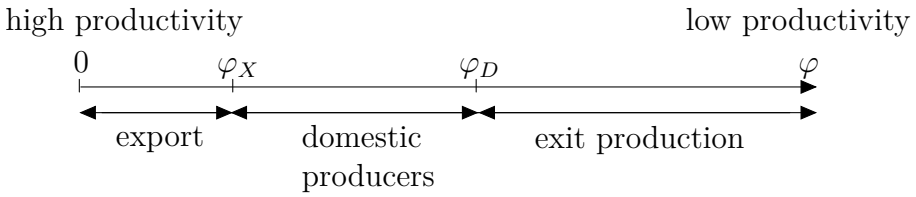


Figure 5.1: Cutoff values and the self-selection

change during the firm's lifetime, then this type of model focuses on the self-selection mechanism, and learning by exporting is not considered.[4]

On the demand side, we follow the standard Dixit-Stiglitz framework which is a simple general equilibrium model with monopolistic goods. Consumers have CES preference with an elasticity of substitution $\epsilon > 1$ across varieties. Solving for the consumer's maximization problem yields the demand for a specific variety of good ω:

$$q(\omega) = Qp(\omega)^{-\epsilon}P^{\epsilon-1},$$

where $P = \left(\int_\Omega p(\omega)^{1-\epsilon}d\omega\right)^{1/(1-\epsilon)}$ is the aggregate price, Q is the exogenous demand level that reflects the market size and $p(\omega)$ is the optimal pricing for the variety ω with a constant mark-up. On the supply side, firms that paid the entry cost draw their productivity from a continuous cumulative distribution $G(\varphi)$, where φ denotes the marginal cost, which is the inverse of productivity. Then, given their productivity, firms decide whether they produce or not, and whether they export or not.

First, firms choose to produce and to serve the domestic market (subscripted by D) if their operating profit is positive. The domestic operating profit is given by $\Pi_D = \Lambda_D\varphi^{1-\epsilon} - f_D$ where $\Lambda_D \equiv Q_D\epsilon^{-\epsilon}(1-\epsilon)^{1-\epsilon}$ represents the domestic demand and f_D is the operating fixed cost of production. Second, firms that export (subscripted by X) can earn an additional profit $\Pi_X = \Lambda_X(\varphi\tau)^{1-\epsilon} - F_X$, where Λ_X represents the foreign demand and $\tau > 1$ denotes the iceberg transport cost between the two countries. F_X represents both the operating fixed cost and the entry cost for the export market. Therefore, three types of firms coexist (see Figure 5.1). Firms that draw a low level of productivity, such as $1/\varphi < 1/\varphi_D$, choose to immediately exit production because their operating profits are negative. Firms that draw a level of productivity between $1/\varphi_D$ and the export cutoff value $1/\varphi_X$, are serving the domestic market only. Firms with a high level of productivity, $1/\varphi > 1/\varphi_X$, serve the domestic market as well as the export market.

[4]Testing effects of learning by exporting is beyond the scope of this chapter. Future research may consider a more dynamic trade model that allows us to study simultaneous the effect of self-selection and of learning by exporting.

5.2.1 Heterogeneous export entry costs

Often, fixed production costs and sunk costs of entry are supposed to be identical for all firms in the literature on heterogeneous firms. These hypotheses seem to be very restrictive regarding the real situation of firms. The differences between two firms in their sunk costs may due to their own specificities or some externalities. Thus, this additional source of heterogeneity is needed to be taken into account. We introduce this heterogeneity through productivity. Our working assumption is that a more productive firm is able to adapt products for foreign markets at a lower entry cost than less productive firms.

As we cannot empirically estimate firms' entry costs in their own domestic market because we do not observe firms that are not producing, these start up costs (f_E) are assumed to be identical across firms. The operating fixed cost for the domestic market f_D is also identical across firms. We focus on the fixed and sunk entry costs for the export market. We define these costs as a function of firm's productivity $F_X(\varphi) \equiv f_X h(\varphi)$, where f_X represents the common sunk cost and $h(\varphi)$ represents firm's deviation from these common sunk costs. We impose that $h(\varphi)$ is an increasing function of φ to satisfy our working assumption that sunk costs are lower for more productive firms. In order to ensure that a firm cannot export without being a domestic producer, we also impose that $f_D < F_X(\varphi)$ for any level of productivity $1/\varphi$. Therefore, the two cutoff values φ_D and φ_X are determined by setting operating profits to zero in each market. The zero profit cutoff conditions for the domestic and export market can be written respectively as:

$$\Lambda_D \varphi_D^{1-\epsilon} = f_D; \tag{5.1}$$

$$\Lambda_X (\tau \varphi_X)^{1-\epsilon} = f_X h(\varphi_X). \tag{5.2}$$

Prior to production, firms have to decide whether they incur or not the start up cost f_E. For a specific firm, the expected operating profit associated to a variety has to be larger than f_E. For the marginal firm, this can be written as:

$$V(\varphi_D)\Lambda_D + \tau^{1-\epsilon} V(\varphi_X)\Lambda_X - \left[G(\varphi_D) f_D + f_X \int_0^{\varphi_X} h(\varphi) dG(\varphi) \right] = f_E, \tag{5.3}$$

where $G(\varphi)$ is the cumulative distribution function from which a potential market entrant draws its productivity and $V(\varphi_j) = \int_0^{\varphi_j} \varphi^{1-\epsilon} dG(\varphi)$ $(j = D, X)$ is the expected value of a firm that draws a productivity φ. In line with Helpman et al. (2004), firms' productivity is drawn from a Pareto distribution, which is a good approximation of data and makes the model more tractable. The Pareto cumulative distribution function is defined as $G(\varphi) \equiv (\varphi/\bar{\varphi})^k$ where φ has a positive support over $[0; \bar{\varphi}]$; $\bar{\varphi}$ and $k > 0$ are the scale and shape parameters, respectively. Since we assume that countries are symmetric, so that the demand shifter is the same in both types of countries, $\Lambda = \Lambda_D = \Lambda_X$. Then,

from (5.1) and (5.2) we can derive the cutoff value ratio to illustrate the self-selection bias of traditional models.

$$\frac{\varphi_X}{\varphi_D} = \left(\frac{f_X}{f_D \tau^{1-\epsilon}}\right)^{\frac{1}{1-\epsilon}} \times h(\varphi_X)^{\frac{1}{1-\epsilon}}, \tag{5.4}$$

where $(f_X/f_D\tau^{1-\epsilon})^{\frac{1}{1-\epsilon}}$ is the cutoff value ratio in the traditional model. Note that our equilibrium ratio in (5.4) is identical to a traditional model ratio except that $h(\varphi)$ now appears on the right-hand side of this expression. The predicted proportion of firms that are between φ_X and φ_D (domestic producers) deviates from the prediction of the traditional model by $h(\varphi_X)^{1/(1-\epsilon)}$.

In order to derive more explicit results, we consider the simplest functional form for $h(\cdot)$ which is compatible with our working assumption, i.e., $h(\varphi) \equiv \varphi^{\gamma}$. The parameter $\gamma \geq 0$ is the productivity elasticity of entry costs. It measures how the change of productivity affects the sunk cost of entry into the export market. If $\gamma = 0$, then $h(\varphi) = 1$ and we are back to the traditional model ($F_X(\varphi) = f_X$). For any $\gamma > 0$, there is a deviation from the common fixed cost. The magnitude of the deviation depends on γ and φ (see section 5.2.2 for a numerical application). For any $\gamma \neq 0$, there is no analytical solution in the present model for φ_D, φ_X and Λ, where the variables are expressed exclusively on the fixed parameters. However, we can compare equilibrium of our model to those of the traditional model by examining the following proposition. From (5.1), (5.2), (5.3) and (5.4) we can obtain the productivity threshold values at equilibrium.[5]

Proposition 5.1. *The unique equilibrium of productivity threshold values is:*

$$\varphi_D^k = \bar{\varphi}^k \frac{f_E(\beta - 1)}{f_D[1 + \Omega(\varphi_X^{\gamma})^{1-\beta}\Delta]}; \tag{5.5}$$

$$\varphi_X^k = \bar{\varphi}^k \frac{f_E(\beta - 1)\Omega}{f_X(\varphi_X^{\gamma\beta} + \Omega\Delta\varphi_X^{\gamma})}, \tag{5.6}$$

where $\beta \equiv k/(\epsilon - 1) > 1$, $\Delta \equiv \beta - (\beta - 1)[k/(k + \gamma)] > 0$ *and* $\Omega \equiv (\tau^{1-\epsilon})^{\beta}(f_X/f_D)^{1-\beta}$. Ω *is a trade openness parameter that is bounded by* 0 *(correspond to autarky) and* 1 *(correspond to free trade).*

The proof of Proposition 5.1 is given in Appendix A. Note that when $\gamma = 0$ our threshold values correspond to the threshold values of the traditional model. From (5.5) and (5.6), we see that φ_X also appears in the right-hand side of the equations whereas it does not in the traditional model. The additional parameter γ affects the equilibrium value of φ_X, φ_D and Λ, which will have an impact on the number of entrants (for both domestic

[5] See Appendix A for details of calculation.

and foreign markets), on the volume of trade and also on consumer welfare.

In this chapter, we focus on what happens to the threshold value, φ_X, at the equilibrium when the productivity elasticity of entry cost, γ, changes. The equilibrium is characterized by an implicit function and it is impossible to solve (5.6) explicitly for φ_X. Therefore, in the next subsection, we carry out the comparative statics by using the Implicit Function Theorem (IFT), and provide a numerical example.

5.2.2 Comparative statics

In our model, the effect of productivity on sunk entry costs is characterized by the elasticity γ. We now investigate analytically the effects of an increase in γ on the equilibrium entry condition, φ_X, for export markets. By fixing all other parameters, Equation (5.6) can be rewritten as the following implicit function of φ_X and γ:

$$F(\varphi_X, \gamma) = \bar{\varphi}^k \frac{f_E(\beta-1)\Omega}{f_X(\varphi_X^{\gamma\beta} + \Omega\Delta\varphi_X^{\gamma})} - \varphi_X^k = 0. \tag{5.7}$$

Given some regularity conditions, the IFT provides the derivative of φ_X w.r.t. γ although it is impossible to solve this implicit function explicitly:

$$\frac{\partial \varphi_X}{\partial \gamma} = -\frac{\frac{\partial F}{\partial \gamma}(\varphi_X, \gamma)}{\frac{\partial F}{\partial \varphi_X}(\varphi_X, \gamma)},$$

where

$$\begin{aligned}\frac{\partial F}{\partial \gamma}(\varphi_X, \gamma) &= -\frac{\bar{\varphi}^k f_E(\beta-1)\Omega}{f_X} \frac{1}{(\varphi_X^{\gamma\beta} + \Omega\Delta\varphi_X^{\gamma})^2} \\ &\quad \left[\beta \log(\varphi_X)\varphi_X^{\gamma\beta} + \Omega \frac{\varphi_X^{\gamma}[\log(\varphi_X)(k+\gamma)(\beta\gamma+k) + (\beta-1)k]}{(k+r)^2}\right];\end{aligned}$$

$$\frac{\partial F}{\partial \varphi_X}(\varphi_X, \gamma) = -\frac{\bar{\varphi}^k f_E(\beta-1)\Omega}{f_X} \frac{(\gamma\beta\varphi_X^{\gamma\beta-1} + \Omega\Delta\gamma\varphi_X^{\gamma-1})}{(\varphi_X^{\gamma\beta} + \Omega\Delta\varphi_X^{\gamma})^2} - k\varphi_X^{k-1}.$$

Using this result, we obtain the following proposition whose proof is provided in Appendix A.

Proposition 5.2. *The three conditions of IFT are satisfied by (5.7): a) $F(.)$ is continuously differentiable function; b) $\partial F(\varphi_X, \gamma)/\partial \varphi_X \neq 0$; c) there is a unique equilibrium.*

i) *For $\varphi_X \in [1, +\infty[$, φ_X is decreasing in γ, i.e., the slope $\partial\varphi_X/\partial\gamma < 0$;*

ii) *For $\varphi_X \in]0, \exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma})]$, φ_X is increasing in γ, i.e., the slope $\partial\varphi_X/\partial\gamma > 0$;*

iii) *For* $\varphi_X \in]\exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma}), 1[$*, the sign of slope is undetermined.*

This proposition shows that there are two schemes. If the export market is relatively open, the *low entry barrier* scheme, i.e., $\varphi_X \in [1, +\infty[$, then Proposition 5.2 indicates that $\partial\varphi_X/\partial\gamma < 0$, implying that the threshold productivity for entry $(1/\varphi_X)$ faced by firms increase as γ increases. Thus, this export market becomes more selective. Conversely, $\partial\varphi_X/\partial\gamma > 0$ when the equilibrium is established in the *high entry barrier* zone, i.e., $\varphi_X \in]0, \exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma})]$, where $\exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma}) < 1$. This suggests that the export market becomes less selective as the parameter γ increases. We note that there is an interval, $\varphi_X \in]\exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma}), 1[$, where the monotonicity is unclear. However, it does not affect our conclusion for two reasons: first, this ambiguous zone is small for a reasonable calibration of the model. Second, it is reducing as γ increases, i.e., $\lim_{\gamma\to+\infty} \exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma}) = 1$.[6]

In order to provide a graphic representation of the proposition above, we solve the model numerically for the variable of interest φ_X. We adopt calibrations as in previous literature on heterogeneous firm models, such as Bernard et al. (2007) and Costantini and Melitz (2007). The parameter calibration is reported in Table 5.1.

Table 5.1: Parameter calibration

Parameter	Value
k	3
ϵ	3.8
f_D	0.1
f_E	2
f_X	0.1 or 2.1
$\bar{\varphi}$	2

Given the calibration of parameters, we are able to plot the equilibrium values of φ_X for different values of γ in Figures 5.2 and 5.3. The low entry barrier scheme is depicted in Figure 5.2, where $\varphi_X \in [1; +\infty[$. In this case, entry becomes more difficult for a higher γ. Figure 5.3 is the relatively selective market situation $\varphi_X \in]0; \exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma})[$, with a high common cost $f_X = 2.1$. In this case, we note that entry is easier for a higher γ. We also see from Figure 5.3 (the red line), that the lower bound of the critical interval $\exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma})$ is very close to one and the equilibrium value of φ_X never falls into this interval. Thus, the slope of φ_X is well determined on its full support in this example.

[6]Following Bernard et al. (2007), we assume that $k = 3$ and $\epsilon = 3.8$. Thus, even when γ is set to be zero, the unclear interval is $]0.976, 1[$.

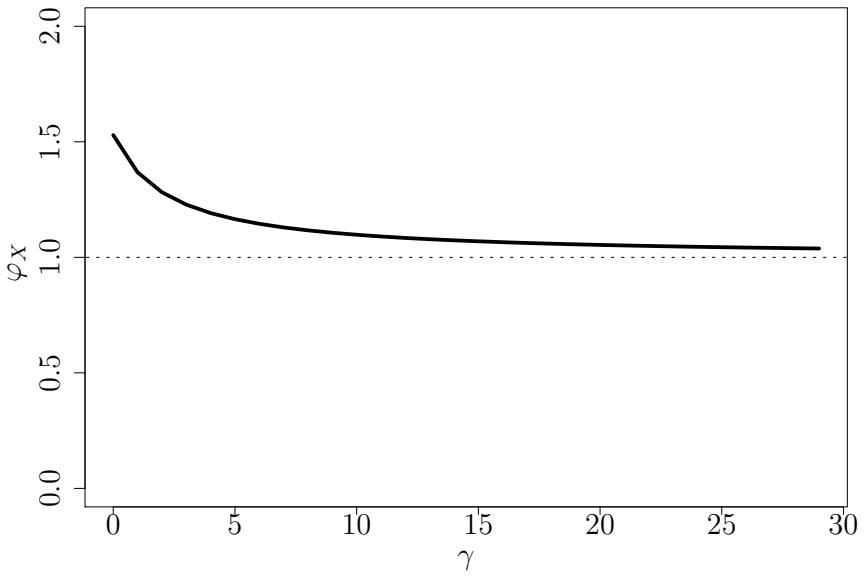


Figure 5.2: Low entry barrier scheme with $f_X = 0.1$

Intuitively, there are two schemes because of the non-monotonicity of $F_X(\varphi) = f_X\varphi^\gamma$ in γ.[7] When the equilibrium value $\varphi_X \in]0, \exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma})[$, the potential entrants with productivity level $\varphi_i < 1$ (in the neighborhood of φ_X), benefit from a higher γ (reduction of $F_X(\varphi)$). In the opposite case, when $\varphi_X \in [1, +\infty[$, the potential entrants with the productivity level $\varphi_i > 1$, suffer from a higher γ (increase in $F_X(\varphi)$). Without loss of generality, we can easily restrict our model to allow for only one of these two situations by bounding the Pareto distribution differently.[8]

In order to validate theoretical implications of heterogeneous fixed costs, we need to estimate sunk entry costs for the export market as well as productivity, and to test the link between them. In particular, we investigate whether and how productivity influences the sunk costs of entry. The answer to these questions is important to evaluate self-selection. As we have shown in this section, the influence of productivity on sunk entry costs can have an effect on the cutoff values (φ_D, φ_X) for market entry. Therefore, the selection process can be stronger (or weaker, depending on the equilibrium schemes) leading to a higher (or lower) aggregate productivity.

[7] When $\varphi < 1$, a higher parameter γ yields a low F_X. Reversely, when $\varphi > 1$, F_X increases with a higher parameter γ.

[8] If the support of the Pareto distribution is $[1; +\infty[$, we only have the low entry barrier scheme.

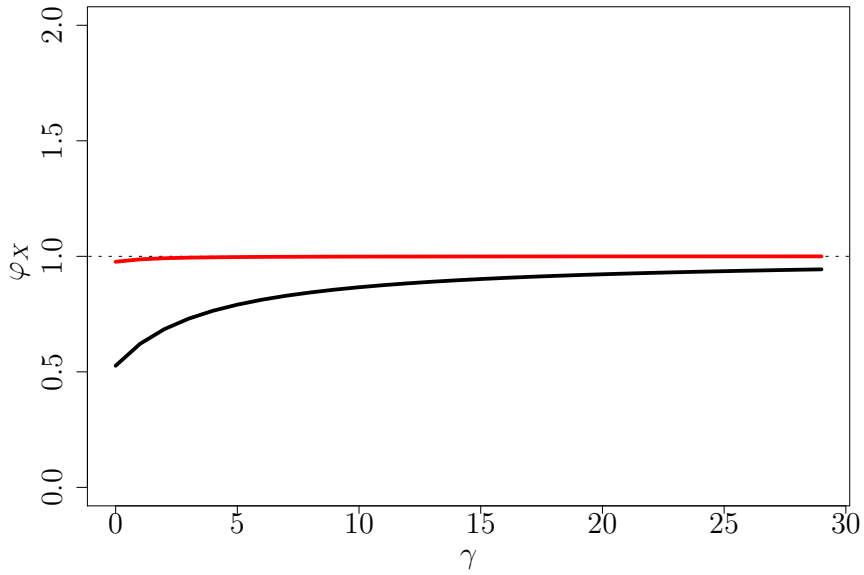


Figure 5.3: High entry barrier scheme with $f_X = 2.1$

5.3 Empirical methodology for measuring entry barriers

A large number of empirical studies, following Roberts and Tybout (1997), investigate the determinants of the export participation decision (for example, Bernard and Wagner, 2001, Campa, 2004, Bernard and Jensen, 2004 and Das et al., 2007). Although firm heterogeneity is often embodied in the trade decision process, the sunk cost of entry is considered to be a common parameter in these empirical studies.[9]

In the basic Melitz (2003) model, firms may have different marginal cost structures (depending on their characteristics and productivity) but all firms face the same sunk cost of entry. Chapter 4 of this thesis, however, shows that fixed and variable costs are not independent and that any heterogeneity affecting variable cost also influences the fixed cost. One of our contributions w.r.t. the previous trade models is that we endogenize the sunk cost of entry as a function of firm-level characteristics. The empirical objective of this chapter is to evaluate the impact of firms' characteristics, in particular productivity, on costs of entry. We propose an empirical strategy that allows us to estimate the firm-specific entry cost and that keeps a minimum level of restrictions on the model.

Following our theoretical model, we disregard the assumption that firms can improve their productivity through exporting. This notion is referred to as *learning-by-exporting*. In this literature, researchers test the effect of entry into exporting markets on firm's

[9] One exception is Das et al. (2007), where the sunk cost of entry depends on firm specific characteristics.

productivity, see Bernard and Jensen (1999), Van Biesebroeck (2005) and De Loecker (2007). In contrast, we study the reverse causality of the productivity-export relationship, which consist of testing whether more productive firms face lower costs of entry into export markets.

In this chapter, we first test for the existence of sunk entry costs. Second, we evaluate the impact of firms' characteristics on their incurred entry costs. In particular, we focus on the impact of neutral and non-neutral productivity. Our empirical model extends the idea of heterogeneous firms in the "new new" trade theory (proposed by Melitz, 2003 and Bernard et al., 2003) by also allowing heterogeneity in terms of entry costs. This empirical investigation supports the exported-oriented policies which favor domestic productivity growth rather than subsidizing entry costs. Such a subsidy policy will not be efficient, first, because entry costs are heterogeneous and the actual expenditure incurred by similar (in terms of export volume) firms can differ significantly from the average level. Second, a large part of these costs are unobservable for the national trade promotion agency.

5.3.1 Overview of our empirical model

The strategy for identifying the sunk cost of entry in our chapter is similar to Roberts and Tybout (1997). This approach consists of comparing the net export profits of newly entered exporters with those of established exporters. However, this chapter differs from their work in the way of estimating sunk costs of entry, and our empirical model allows for entry cost heterogeneity.

The logarithmic net export profits for exporter i at period t ($\pi_{it} \equiv \log\Pi_{it}$) can be written as:

$$\pi_{it} = \begin{cases} \pi_{0it} = \pi^*_{it} + v_{0it} & \text{if } D_{it-1} = 0; \\ \pi_{1it} = \pi^*_{it} - \log f_X + v_{1it} & \text{if } D_{it-1} = 1. \end{cases} \tag{5.8}$$

The net current profit depends on firm's previous exporting status, where we distinguish two cases that are denoted by π_{0it} and π_{1it}. The binary variable $D_{it-1} = 0$ indicates that the exporter is "on the export market" at period $t-1$ (so the *established exporters* since $t-1$) and $D_{it-1} = 1$ indicates that the exporter is "out of the export market" at period $t-1$ (as the *newly entered exporter* at t). The logarithmic gross exporting profit is denoted as π^*_{it}. The terms v_{0it} and v_{1it} represent the individual-specific gain (or loss) from export participation, which are anticipated by firms but unobserved by econometricians.

The difference between the i-th newly entered exporter's net profits (observed) and his potential profits without incurring the (unobserved) entry cost defines the sunk cost of entry, i.e., $\log F_{Xit} = \pi_{0it} - \pi_{1it} = \log f_X + v_{0it} - v_{1it}$. If we only had data on newly entered exporters (with $D_{it-1} = 1$), it would be infeasible to separately identify the sunk cost of entry from the expected profits. Fortunately, it is generally possible to

have a control group of established exporters (with $D_{it-1} = 0$), which are "already on the export market" at $t-1$. The sunk cost is identified as the difference in terms of actual profits between the newly entered exporters and the established exporters which have comparable potential profit streams.

An additional empirical difficulty in this framework is that our data set does not provide information on the export profit, but we observe instead the export revenue ($r_{it} \equiv \log R_{it}$). We follow Das et al. (2007), and assume the classical markup equation:

$$p_{it}(1 - \eta_i^{-1}) = c_{it},$$

where p_{it} denotes the foreign output price, c_{it} is the marginal cost for the export production, and $\eta_i > 1$ denotes the firm-specific foreign markup. Multiplying both sides of the markup equation by the export quantities yields:

$$R_{it}(1 - \eta_i^{-1}) = C_{it},$$

where C_{it} is the variable cost of exporting. Then we can express the export profit as:

$$\Pi_{it} = R_{it} - C_{it} = \eta_i^{-1} R_{it}.$$

Substituting the equation above into model (5.8), we obtain the corresponding empirical model in terms of revenues:

$$r_{it} = \log\eta_i + \pi_{it} = \begin{cases} r_{0it} = \log\eta_i + \pi^*_{it} + v_{0it} & \text{if } D_{it-1} = 0 \\ r_{1it} = \log\eta_i + \pi^*_{it} - \log f_X + v_{1it} & \text{if } D_{it-1} = 1. \end{cases} \tag{5.9}$$

In the earlier empirical literature, researchers estimated the sunk cost as an average constant term by considering that the sunk cost is identical across firms. Several attempts to incorporate the firm-level heterogeneity have focused on the heterogeneity in the trade decision process rather than on the sunk cost structure, see for example Bernard and Jensen (2004). In our theoretical model, the sunk cost of entry is not constant but a function of firm's characteristics. To capture the feature of endogenous sunk costs, our empirical model allows firms to deviate from the average entry requirement. The average sunk cost of entry ($\log f_X$) is estimated along with the firm's deviation ($v_{0it} - v_{1it}$) from the average sunk cost. Compared to the dynamic discrete-choice model proposed by Roberts and Tybout (1997), our estimation method follows the treatment evaluation literature (Heckman and Hotz, 1989), and matches newcomers with established exporters which share similar characteristics. The structural model proposed by Das et al. (2007) can also be used to characterize the heterogeneity in the sunk cost of entry. However, we prefer the treatment evaluation approach because it avoids to impose further *ad hoc* assumptions on the functional form and on the dynamic

stochastic process for expected profits and sunk costs.

The matching method compares the current export revenue of the two groups (the treated with the control group) based on a series of criteria which are embodied in a vector of pretreatment variables. These variables reflect the firm's financial and production information as well as its productivity. An additional novelty of this chapter is that we consider two types of productivity: Hick-neutral (Total Factor Productivity, TFP) and non-neutral (Relative Factor-augmenting Productivity, RFP). Following Chapter 3, we estimate two measurements of productivity and evaluate their impacts on firm-specific entry costs. In the next subsections, we first present the treatment evaluation model for testing the existence of sunk entry costs and for quantifying the impact of firm-level characteristics on the sunk cost of entry. Then, we describe the estimation method for productivity measurement.

5.3.2 Treatment evaluation model

We formally define the sunk cost of entry for export markets:

Definition 5.1. *At period t, the sunk cost of entry for* i*-th newly entered firm is defined as the difference between the actual (observed) export revenue (r_{1it}) and its potential (unobserved) export revenue without incurring the entry cost (r_{0it}):*

$$\begin{aligned} \log F_{Xit} &= r_{0it} - r_{1it} = \log f_X + (v_{0it} - v_{1it}) \\ &= \theta + (v_{0it} - v_{1it}), \end{aligned} \tag{5.10}$$

where $\theta \equiv \log f_X$ represents the average sunk cost of entry into the exporting market. The term $(v_{0it} - v_{1it})$ that is time-varying and firm-specific, represents the individual deviation from the average sunk cost.

If the same firm could be observed in both states at the same time period, the sunk cost of entry could simply be calculated. However, the difficulty is that a firm cannot be in both states simultaneously. We observe either r_{0it} or r_{1it} for the i-th individual at t. Basically, we are facing a missing data problem. To overcome this problem, we use the matching technique to estimate the effect of entry. Our group of interest (the treated group) includes newly entered exporters with $D_{it-1} = 1$ that earn r_{1it} in the export market at t. The control group includes established exporters with $D_{jt-1} = 0$ that earn r_{0jt}, with $i \neq j$. Intuitively, the sunk cost of entry is revealed by comparing the difference in the actual exporting revenue at t across exporters that have comparable expected profits and characteristics, but differ in whether they exported in the previous period. The underlying hypothesis here is that sunk costs of entry are borne completely within one year after the entry.

Our empirical objective is not to obtain the structural parameters as in Das et al. (2007), but to evaluate the cost of entry and to study how this cost changes with some firm's characteristics. Therefore, we avoid *ad hoc* restrictions on the expect profit function and the markup. The quantity of interest is the average treatment effects (ATE) conditional on firm's *pretreatment characteristics* x_{it-1}:

$$ATE(x_{it-1}) = \mathrm{E}[r_1 - r_0 \mid x_{it-1}]. \tag{5.11}$$

The vector of covariates x, may include firm's past productivity and size. Equation (5.11) measures the average cost of entry cost in terms of export revenue shortfalls. It also characterizes how the cost of entry changes for various level of x, in particular, we are interested in the impacts of TFP and RFP on entry cost into the export market.

The difficulty in estimating ATE is that the treatment (the previous exporting experience, D_{it-1}) is certainly not randomized across firms, and the firm's decision variable at $t-1$, D_{it-1}, may be related to its anticipation of the entry that is captured in v_{0it} and v_{1it}. To deal with this problem, several approaches are available in the literature. Wooldridge, (2002, p.602), Heckman and Vytlacil (2007) and Imbens and Wooldridge (2009) provide recent reviews on this rapidly growing literature. We use the Heckman and Hotz (1989) approach to estimate the treatment evaluation model. A fundamental assumption required by this approach is the *ignorability of treatment* assumption (Rosenbaum and Rubin, 1983).

Assumption 5.1. (Ignorability of treatment) *Conditional on pretreatment characteristics x_{it-1}, the decision variable D_{it-1} and the outcomes (r_{0it},r_{1it}) are mean independent.*

$$\mathrm{E}[r_{0it} \mid D_{it-1}, x_{it-1}] = \mathrm{E}[r_{0it} \mid x_{it-1}];$$

$$\mathrm{E}[r_{1it} \mid D_{it-1}, x_{it-1}] = \mathrm{E}[r_{1it} \mid x_{it-1}].$$

In order to see the implication of this assumption, we recall the export revenue equations (5.9), which can be viewed as a switching regression model:

$$\begin{aligned} r_{it} &= (1 - D_{it-1})r_{0it} + D_{it-1}r_{1it} \\ &= r_{0it} + D_{it-1}(r_{1it} - r_{0it}) \\ &= \log\eta_i + \pi^*_{it} + v_{0it} - \theta D_{it-1} + D_{it-1}(v_{1it} - v_{0it}). \end{aligned}$$

Endogeneity arises because the previous entry decision (D_{it-1}) is related to the firm's expectation about entry benefits or costs, which is reflected in the unobserved term ($v_{1it} - v_{0it}$). Under the ignorability of treatment assumption, the conditional expectation

of outcome becomes:

$$\mathrm{E}[r_{it} \mid D_{it-1}, x_{it-1}] = f(x_{it-1}) + g_0(x_{it-1}) - \theta D_{it-1} + D_{it-1}[g_1(x_{it-1}) - g_0(x_{it-1})],$$

where $g_0(x_{it-1}) \equiv \mathrm{E}[v_{0it} \mid D_{it-1}, x_{it-1}] = \mathrm{E}[v_{0it} \mid x_{it-1}]$ and $g_1(x_{it-1}) \equiv \mathrm{E}[v_{1it} \mid D_{it-1}, x_{it-1}] = \mathrm{E}[v_{1it} \mid x_{it-1}]$. Both expected profits and the markup are a function of firm's characteristics, i.e., $f(x_{it-1}) \equiv \mathrm{E}[\log\eta_i + \pi^*_{it} \mid x_{it-1}]$. By rearranging the terms, the equation above yields:

$$\mathrm{E}[r_{it} \mid D_{it-1}, x_{it-1}] = G_0(x_{it-1}) - \theta D_{it-1} + D_{it-1}G_1(x_{it-1}),$$

where $G_0(x_{it-1}) \equiv f(x_{it-1}) + g_0(x_{it-1})$ and $G_1(x_{it-1}) \equiv g_1(x_{it-1}) - g_0(x_{it-1})$. From the above equations, we note that the dependence between D_{it-1} and the unobserved terms, v_{0it} and v_{1it} are eliminated by conditioning on x_{it-1}. The conditional ATE in this model is:

$$ATE(x_{it-1}) = -\theta + G_1(x_{it-1}).$$

Under the ignorability of treatment assumption, the two quantities of interest can be estimated in a fairly flexible way without imposing any distributional restrictions on the observed outcome. However, in order to simplifying the estimation, we consider a linear model where the parameters of interest are obtained by regressing r_{it} on x_{it-1}, D_{it-1} and $D_{it-1}(x_{it-1} - \bar{x})$, where $\bar{x}$ denotes the sample average of x_{it-1}:

$$r_{it} = \lambda_1' x_{it-1} - \theta D_{it-1} + \lambda_2' D_{it-1} \cdot (x_{it-1} - \bar{x}) + e_{it}. \tag{5.12}$$

The error term e_{it} is assumed to be i.i.d. Consistent estimates of parameters in (5.12) yield the prediction: $\widehat{ATE}(x_{it-1}) = -\hat{\theta} + \hat{\lambda}_2' D_{it-1} \cdot (x_{it-1} - \bar{x})$. $\hat{\theta}$ measures the cost of entry for the newcomer in comparison to the established exporter. If the estimated $\hat{\theta}$ is significant, it suggests that there is a potential entry cost in the export market. A consistent estimate of $\widehat{ATE}(x_{it-1})$ also allows us to evaluate how the ATE given x_{it-1} changes with a particular element of x_{it-1}. The estimated vector of parameter $\hat{\lambda}_2$ indicates the impact of pretreatment variables on the entry effect. If $\hat{\theta}$ is significant (there is a sunk costs of entry), a positive value of $\hat{\lambda}_2$ suggests that x_{it-1} can contribute to reducing the sunk cost of entry.

A more practical question now is: what are the suitable pretreatment variables in x_{it-1}? Given the data at hand, we include firm's characteristics that may affect the decision of entry into the export market and past performances: firm's size, core business sector, as well as the two measurements of productivity at period $t-1$.[10] In the next subsection, we describe how Hicks-neutral and non-neutral productivity (TFP

[10] x_{it-1} includes only the pretreatment variables. Thus, we assume that firms chose to enter into export markets at t after knowing their size and productivity at $t-1$.

and RFP) are measured.

5.3.3 Estimating the productivity

There is a long tradition of studying the productivity-export relationship in the literature. However, in the majority of cases (Crozet and Trionfetti, 2011 is an exception), productivity is referred to TFP (Hicks-neutral). Syverson (2011) and Van Beveren (2012) provide recent surveys on this literature. We take into account a second type of productivity measurement, RFP (non-neutral), for analyzing the firm's trade behavior.

Chapter 3 of this thesis proposes a structural semi-parametric estimation method for recovering the firm-level productivity. This approach extends the Olley and Pakes (1996) estimator to the more flexible and realistic specification of CES production function with biased technical change. The advantage of this approach is that it not only estimates TFP, but also yields time-varying and firm-specific estimates of RFP without prior assumption on its functional form. This approach deals with two sources of endogeneity: the regressors in the production function (input variables) may correlate with both Hicks-neutral and non-neutral productivity. By using the first order conditions derived from competitive factor market, this method allows consistent estimation of the degree of returns to scale, the elasticity of substitution, and the bias in technical change.

The CES production function with two factors, labor (L_{it}) and capital stock (K_{it}), and the value-added output, Y_{it} can be written as:

$$Y_{it} = A_{it}[\alpha(B_{it}K_{it})^{\frac{\sigma-1}{\sigma}} + (1-\alpha)L_{it}{}^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma\rho}{\sigma-1}}, \tag{5.13}$$

where the parameters α, σ and ρ are the income distribution parameter, the degree of returns to scale and the elasticity of substitution, respectively. A_{it} is the relative Hicks-neutral productivity (TFP), and B_{it} is the relative capital-augmenting productivity (RFP), see Chapter 3 for more details on this point. We assume that the productivity term A_{it} follows a first-order Markov process. Cost minimizing firms set the marginal product equals to input prices. The first order conditions of the CES production function imply that:

$$\frac{K_{it}}{L_{it}} = \left(\frac{\alpha}{1-\alpha}\right)^{\sigma}\left(\frac{W_{lit}}{W_{kit}}\right)^{\sigma} B_{it}^{\sigma-1}, \tag{5.14}$$

where W_l and W_k denote the wage and the capital rental price, respectively. The logarithmic CES production function can be rewritten as:

$$\log Y_{it} = \rho \log L_{it} + \frac{\sigma\rho}{\sigma-1}\log[\alpha(B_{it}\frac{K_{it}}{L_{it}})^{\frac{\sigma-1}{\sigma}} + (1-\alpha)] + \log A_{it},$$

and the capital-labor ratio equation (5.14) yields:

$$\log[\alpha(B_{it}\frac{K_{it}}{L_{it}})^{\frac{\sigma-1}{\sigma}} + (1-\alpha)] = \log(1-\alpha) + \log S_{it}, \tag{5.15}$$

where $S_{it} \equiv \frac{W_{kit}K_{it}}{W_{lit}L_{it}} + 1$ reflects the factor cost ratio.[11] Then, we can use (5.15) to substitute the unobservable productivity shock B_{it} from the production function and to obtain the following regression equation:

$$y_{it} = c + \rho l_{it} + \delta s_{it} + a_{it} + \varepsilon_{it}. \tag{5.16}$$

The random term ε_{it} corresponds to the exogenous shock that is not anticipated by firms. The lower-cases denote the logarithmic values. The parameters c and δ are defined as $c \equiv \frac{\rho\sigma}{\sigma-1}\log(1-\alpha)$ and $\delta \equiv \frac{\rho\sigma}{\sigma-1}$.

Now, we have a log-linear model, in which we need to deal with two sources of endogeneity through $a_{it} \equiv \log A_{it}$ and $b_{it} \equiv \log B_{it}$ (both a_{it} and b_{it} are correlated with regressors l_{it} and s_{it}). The seminal paper of Olley and Pakes (1996) introduces the so-called control function approach to deal with the endogeneity problem for estimating production functions. The idea behind this approach is to use a control function of proxies for inverting out the unobserved productivity term in the production function. Levinsohn and Petrin (2003) assume the monotonicity of material demand equation in a_{it}, and inverse a_{it} out from the material demand equation. Thus, the productivity term a_{it} is expressed as a function of the material demand and capital stocks, see Chapter 3 for details. Note that Olley and Pakes (1996) as well as Levinsohn and Petrin (2003) do not allow for biased technical change in their model, because of the choice of a Cobb-Douglas specification. Based on the CES model, we consider a generalized material demand equation, $m_{it} = M(a_{it}, b_{it}, k_{it})$ where both terms a_{it} and b_{it} are included in the equation.[12] One technical difficulty of this generalization is that the inversion trick in Levinsohn and Petrin (2003) is not longer working for a_{it}. The monotonicity of the material demand equation in a_{it} is not sufficient for the inversion of the material demand equation, because the additional term b_{it} is unobserved. Fortunately, using the capital-labor ratio equation (5.14), b_{it} can be expressed as a function of the observed variable S_{it} and the input price ratio.[13] Thus, we can obtain the following generalized

[11] Given the capital-labor ratio equation (5.14),

$$\begin{aligned} & W_{kit}K_{it}/W_{lit}L_{it} = \alpha/(L_{it}/K_{it})^{1/\sigma-1}B_{it}^{(\sigma-1)/\sigma} \\ \iff \quad & (1-\alpha)(W_{kit}K_{it}/W_{lit}L_{it}+1) = \alpha(\tfrac{L_{it}}{K_{it}}B_{it})^{(\sigma-1)/\sigma} + (1-\alpha) \end{aligned}$$

[12] The econometric model considered here differs slightly from Chapter 3 in two aspects: first, we follows Levinsohn and Petrin (2003) by using material input as proxy. Second, the control function now includes the additional term, b_{it}, which allows for the interaction between inputs demands and RFP. Therefore, the corresponding estimation method is also modified.

[13] Inverting the capital-labor ratio equation yields:

$$B_{it} = \left(\frac{1-\alpha}{\alpha}\right)^{\frac{\sigma}{\sigma-1}} (S_{it}-1)^{\frac{1}{\sigma-1}} \left(\frac{W_{kit}}{W_{lit}}\right)$$

invertible relationship.[14]

$$\mathrm{E}[a_{it} \mid s_{it}, k_{it}, m_{it}, w_{kit}, w_{lit}] = M^{-1}(s_{it}, k_{it}, m_{it}, w_{kit}, w_{lit}).$$

Plugging this function into Equation (5.16), we have our final regression equation:

$$\begin{aligned} y_{it} &= c + \rho l_{it} + \delta s_{it} + M^{-1}(s_{it}, k_{it}, m_{it}, w_{kit}, w_{lit}) + \varepsilon_{it} \\ &= \rho l_{it} + \Phi(s_{it}, k_{it}, m_{it}, w_{kit}, w_{lit}) + \varepsilon_{it}, \end{aligned}$$

which is a partial linear model. It is clear that the parameters c and δ are not identified separately from the nonparametric function. Thus, only the parameter ρ and the nonparametric function $\Phi(.)$ are estimated in the first-stage using Robinson's (1988) estimator. Then, given a candidate value of δ, the estimates of a_{it} can be expressed as: $\hat{a}_{it}(\delta) = \hat{\Phi} - \delta s_{it}$.

In the second stage, we need at least one additional moment condition for the estimation of δ. For this purpose, we impose the first-order Markov assumption on a_{it}, and decompose the current TFP as:

$$a_{it} = \mathrm{E}[a_{it} \mid a_{it-1}] + \xi_{it}.$$

The term ξ_{it} is the innovation shock, which represents a deviation of a_{it} from its expectation at $t-1$. We assume that the current composition of factor cost (s_{it}) is chosen by firms at $t-1$.[15] As consequence, the second-stage moment condition is: $\mathrm{E}[\xi_{it} \cdot s_{it}] = 0$. Given the first-stage estimation, the innovation shock ξ_{it} can be written as: $\hat{\xi}_{it}(\delta) = \hat{a}_{it}(\delta) - \mathrm{E}[\hat{a}_{it}(\delta) \mid \hat{a}_{it-1}(\delta)]$. We estimate the parameter δ by minimizing the sample analogues of $\mathrm{E}[\hat{\xi}_{it}(\delta) \cdot s_{it}] = 0$.

Given the estimates of ρ and δ, we can recover an estimate of TFP up to a constant term as:

$$\widehat{TFP} = \hat{a}_{it} + c = y_{it} - \hat{\rho} l_{it} - \hat{\delta} s_{it}. \tag{5.17}$$

Given the implicit estimates of $\sigma = \frac{\rho}{\rho - \delta}$ and the capital-labor ratio equation (5.14), an estimate of RFP up to a constant term (d) is:

$$\widehat{RFP} = -\hat{b}_{it} + d = \frac{1}{(1-\hat{\sigma})} \log(\frac{K_{it}}{L_{it}}) - \log(\frac{W_{kit}}{W_{lit}}). \tag{5.18}$$

This estimate reflects the relative labor-augmenting productivity. In the next sections,

[14] The corresponding production timing assumption is that the material demand is fully flexible (and monotone in a_{it}), which is decided after knowing the capital stock (k_{it}) and the factor cost ratio (s_{it}).

[15] This assumption can be justified by the fact that labor is quasi-fixed. Typically, the capital-labor cost composition (s_{it}) is chosen prior to t, if there is a training process before a worker can actually enter production at t, or if there is a significant hiring cost. More discussion about the production timing can be found in Ackerberg et al. (2006).

we apply our empirical strategy to French manufacturing data.

5.4 Empirical investigation

For our empirical investigation, we use two different sources of data. The Amadeus database provides us with production and financial information for more than 7000 large and very large French manufacturing firms over the period of 2003-2009.[16] The variables included in our data set are operating revenue, export revenue, sales, capital stock, number and costs of employees and costs of material. We merge this information with the price index constructed using INSEE's sector-level series in order to deflate revenues and intermediate material inputs. Our data set at hand allows us to estimate productivity, sunk costs of entry as well as the effects of different factors on the entry cost.

Table 5.2 displays the macroeconomic conditions for trade in France and the distribution of exporting status across firms in the sample. We compare the export participation rates of large and very large firms with the macroeconomic trade index, i.e., the export volume of manufactured goods in France (with the base-year in 2005) and the exchange rate index of the U.S. dollar against the Euro (with the base-year in 2005). Despite changes in the foreign trade environment for French manufacturing firms over the sample period (for instance, the exchange rate of the U.S. dollar against the Euro decrease from 110.1 to 89.4), we observe from Table 5.2 that the proportion of exporters is relatively stable in average this proportion is 76%. Typically, this can be explained by the fact that the sunk cost of entry into the export market produces hysteresis in firm's trade behavior (Baldwin and Krugman, 1989). The entry and exit (in the export market) transition matrix for each pair of two years are reported in Table 5.3. We see that export market entry and exit dynamics are rather stable over the period of 2003-2009.

Our empirical investigation focuses on the entry of firms into export markets. Based on this data set, our estimation method follows the treatment evaluation literature that matches pairs of newly entered and established exporters which share the similar characteristics. For each panel of two years, the control group includes the established exporters and the treated group includes the newly entered exporters. For example, in the period from 2003 to 2004, the control group includes the 3234 firms that are already on the export market in 2003 and continue to serve the foreign market in 2004. The treated group are the 220 newly entered firms in 2004. In this study we consider only pairs of two years, mainly because it is the case where we have a significant number of

[16]The original data set includes 8196 firms, only 7140 French firms (1056 "Credit needed" firms are excluded) for the period 2003-2009 are downloaded from Amadeus database. The definitions of large and very large firms: *very large* firms are defined as the firms that *Turnover* $>$ 100 million Euros or *Total assets* $>$ 200 million euros or *Employees* $>$ 1000. *Large* firms are defined as the firms that *Turnover* $>$ 10 million Euros, or *Total assets* $>$ 20 million euros, or *Employees* $>$ 150.

Table 5.2: Export participation of French manufacturing firms 2003-2009

	2003	2004	2005	2006	2007	2008	2009
Exchange rate index (1)	110.1	100.1	100	99.1	90.8	84.9	89.4
Export volume index (2)	93.1	97.2	100	107.9	109.1	107.8	93.8
Exporters in % (3)	76.5	76.4	77.2	76.4	76.0	75.0	74.0

Souce: INSEE

Note: (1) - Exchange rate index is the annual average exchange rate of of U.S. dollar (for 1 dollar) against Euro with base year in 2005. (2) - Export volume index is the French aggregate export volume index with base year in 2005 for all manufactured goods. (3) - Exporters in % indicates the percentage of exporters in the sample.

Table 5.3: Transition rates in the export market 2003-2009

	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09
Outsider (1)	616	616	628	603	549	477
Insider (2)	3234	3254	3141	2996	2561	2120
Entry	220	175	153	175	131	120
Exit	161	171	189	167	174	149
Total firms (3)	4231	4216	4111	3941	3415	2866
Outsider in %	14.6	14.6	15.3	15.3	16.1	16.6
Insider in %	76.4	77.2	76.4	76.0	75.0	74.0
Entry in %	5.2	4.2	3.7	4.4	3.8	4.2
Exit in %	3.8	4.1	4.6	4.2	5.1	5.2

Note: (1) - *Outsider* indicates the number of non-exporting firms. (2) - *Insider* indicates the number of exporters. (3) - Total firms indicates the total number of firms in the sample, whose production and trade information are available in both years.

new entrants and controls.

5.4.1 Estimation of productivity

Following the methodology described in Section 5.3.3, we estimate both the Cobb-Douglas and the CES production function. The method proposed in Chapter 3 is based on the CES production function and the Levinsohn and Petrin (2003) method is based on the Cobb-Douglas specification. More details on the second method can be found in Levinsohn and Petrin (2003).

Given the estimates of parameters, TFP and RFP are recovered by using (5.17) and (5.18). The estimates of productivity are used in the second stage for the matching model, and the matching consists of comparing firms with comparable characteristics at period $t-1$. Therefore, we need to find the comparable productivity measurement between the treated and control groups. The problem is that, at period $t-1$, the control group (including established firms) serves both foreign and domestic markets while the treated group (including newly entered firms) serves only the domestic market. Therefore, we match the two groups with their domestic market productivity at $t-1$. Formally, in order to measure only the domestic market productivity, we use the domestic value-added as the dependent variable in the production function regression.[17]

Table 5.4 presents estimates of technology parameters. β_l and β_k denote the coefficients of labor and capital in the Cobb-Douglas production function. ρ, δ and σ are the technology parameters of the CES specification (5.13). The value of the elasticity of substitution σ is the fundamental difference between the Cobb-Douglas and CES production functions. This elasticity is restricted to be one in the Cobb-Douglas specification while it is a free parameter in the CES case. From the estimation results of Table 5.4, we note that the estimated elasticity of substitution is significantly below one. This finding rejects the use of Cobb-Douglas specification in favor of the CES model and suggests that capital and labor are complements for production. Chapter 3 points out that, compared to the CES model, the Cobb-Douglas model yields a lower estimate of returns to scale and this bias is essentially due to the omitted-variable-bias. In our data, we confirm the prediction of Chapter 3 (Table 3.1) that the estimates of returns to scale in the Cobb-Douglas case is lower than the estimates of returns to scale in the CES case.

Given the estimates of technology parameters, we compute the correlation coefficients between estimated domestic productivity, domestic output, labor, capital stock and the export revenue. The correlation matrix for 2003-2004 is reported in Table 5.5. Table 5.8 and Figure 5.4 in Appendix B provide the correlation matrix in other periods and the distributions of estimated productivity, respectively. The two estimated TFP

[17]The domestic intermediate materials is calculated as the total intermediate materials weighted by domestic production share. Thus, the domestic value-added is defined as gross domestic output net of domestic intermediate materials.

Table 5.5: Correlation matrix between productivity and firm's characteristics

2003-2004	$\widehat{TFP}^{CB}_{t-1}$	$\widehat{TFP}^{CES}_{t-1}$	$\widehat{RFP}^{CES}_{t-1}$	Y_{t-1}	L_{t-1}	K_{t-1}
$\widehat{TFP}^{CES}_{t-1}$	0.94					
$\widehat{RFP}^{CES}_{t-1}$	0.07	0.31				
Y_{t-1}	0.83	0.73	0.17			
L_{t-1}	0.39	0.18	0.03	0.73		
K_{t-1}	0.33	0.36	0.71	0.64	0.72	
r_t	0.06	-0.01	0.18	0.34	0.53	0.51

Note: the subscript "t" and "$t-1$" indicate the 2004 and 2003 variables, respectively.

that are obtained from the Cobb-Douglas and CES models, are highly correlated and have similar distribution shapes. The CES model provides an additional productivity measurement, RFP. The estimate of RFP is positively correlated with TFP and its distribution is flatter. We note that RFP is highly correlated with capital stock. This may reflect the fact that the labor-augmenting innovation is mainly realized through investment in capital. The RFP is also positively correlated with the next period exporting revenue.

Table 5.4: Estimation of technology parameters

	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09
Cobb-Douglas						
β_l	0.511 (0.024)	0.510 (0.021)	0.468 (0.024)	0.492 (0.021)	0.493 (0.025)	0.511 (0.029)
β_k	0.080 (0.030)	0.068 (0.040)	0.006 (0.030)	0.077 (0.041)	0.067 (0.057)	0.304 (0.120)
$\beta_k + \beta_l$	0.591	0.578	0.474	0.569	0.560	0.815
CES						
ρ	0.736 (0.034)	0.731 (0.036)	0.615 (0.037)	0.589 (0.039)	0.607 (0.051)	0.780 (0.053)
δ	−0.319 (0.172)	−0.440 (0.169)	−0.579 (0.141)	−0.405 (0.122)	−0.386 (0.131)	−0.332 (0.257)
σ	0.302 (0.133)	0.376 (0.095)	0.485 (0.071)	0.407 (0.081)	0.389 (0.091)	0.298 (0.230)

Note: the bootstrap estimated standards deviations are reported in parenthesis.

5.4.2 Entry costs and productivity

In our matching model, the treated group are firms which become exporters at period t and serve only the domestic market at period $t-1$. The control group consists of established exporters which serve both the foreign and the domestic market for t and $t-1$. The treatment indicator variables D is defined as $D = 1$ for the individuals in

Table 5.6: Estimation of $ATE(x)$

	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09
Intercept	6.27 (4.00)	4.40 (4.66)	7.30 (2.44)	7.99 (3.00)	2.10 (3.84)	−3.41 (3.11)
D_{it-1}	−1.12 (0.18)	−2.07 (0.26)	−1.26 (0.33)	−2.08 (0.34)	−1.51 (0.27)	−2.01 (0.33)
D_{it-1}·**TFP**$_{it-1}$	1.05 (0.35)	2.28 (0.37)	0.66 (0.56)	2.09 (0.59)	1.35 (0.46)	2.35 (0.56)
D_{it-1}·**RFP**$_{it-1}$	−0.09 (0.11)	−0.32 (0.15)	0.03 (0.15)	−0.19 (0.14)	−0.14 (0.16)	−0.15 (0.19)
D_{it-1}·Sector$_i$	$9.57e^{-4}$ ($2.35e^{-4}$)	$7.96e^{-4}$ ($3.56e^{-4}$)	$2.28e^{-4}$ ($3.12e^{-4}$)	$1.28e^{-3}$ ($3.85e^{-4}$)	$1.90e^{-4}$ ($3.95e^{-4}$)	$9.65e^{-4}$ ($3.70e^{-4}$)
D_{it-1}·Labor$_{it-1}$	0.89 (0.58)	1.44 (0.99)	1.96 (1.10)	−0.16 (1.00)	0.57 (0.83)	−0.17 (1.12)
D_{it-1}·Squared labor$_{it-1}$	−0.09 (0.05)	−0.20 (0.10)	−0.18 (0.12)	−0.03 (0.11)	−0.06 (0.08)	−0.02 (0.10)
D_{it-1}·Wage$_{it-1}$	6.63 (4.93)	−15.10 (7.84)	−1.77 (9.45)	−12.19 (6.19)	−2.17 (7.01)	−8.68 (7.57)
D_{it-1}·Squared wage$_{it-1}$	−0.89 (0.64)	1.81 (1.03)	0.21 (1.29)	1.49 (0.86)	0.19 (0.90)	0.82 (1.00)
TFP$_{it-1}$	−0.91 (0.05)	−0.87 (0.05)	−0.81 (0.05)	−0.83 (0.05)	−0.86 (0.06)	−0.70 (0.06)
RFP$_{it-1}$	0.27 (0.03)	0.25 (0.03)	0.20 (0.02)	0.23 (0.03)	0.22 (0.03)	0.20 (0.03)
Sector$_i$	$-1.82e^{-5}$ ($5.23e^{-5}$)	$-5.00e^{-6}$ ($5.21e^{-5}$)	$5.51e^{-5}$ ($5.63e^{-5}$)	$5.48e^{-5}$ ($5.56e^{-5}$)	$3.73e^{-5}$ ($5.80e^{-5}$)	$4.59e^{-6}$ ($6.25e^{-5}$)
Labor$_{it-1}$	0.35 (0.20)	−0.02 (0.20)	−0.19 (0.25)	−0.19 (0.28)	−0.14 (0.23)	−0.19 (0.25)
Squared labor$_{it-1}$	0.09 (0.02)	0.12 (0.02)	0.14 (0.02)	0.14 (0.03)	0.13 (0.02)	0.12 (0.02)
Wage$_{it-1}$	−1.12 (2.13)	0.34 (2.62)	−0.73 (1.31)	−0.92 (1.58)	2.09 (1.96)	4.27 (1.59)
Squared wage$_{it-1}$	0.51 (0.29)	0.31 (0.37)	0.43 (0.18)	0.44 (0.21)	0.06 (0.30)	−0.22 (0.20)

the treated sample and $D = 0$ for the individuals in the control sample. The ATE and $ATE(x)$ are estimated using the regression based method (see Heckman and Hotz, 1989).

In our specification of the empirical model (5.12), the regressors vector x_{it-1} includes an indicator for the sector of activity ($Sector$), labor (l), wage (w_l), squared labor (l^2), squared wage (w_l^2) and the estimated productivity $\widehat{TFP}$ and $\widehat{RFP}$.[18] Table 5.6 summarizes the estimation results as well as the estimated standard deviations (reported in parentheses). Since the estimates are obtained by using a step-wise estimation approach (a first-stage yields TFP and RFP), we estimate standard deviations by panel bootstrapping. A robustness check is reported in Appendix B, where the treatment evaluation model is re-estimated using a propensity score matching.

We note that the estimate of $-\theta$ (the estimated coefficients of D_{it-1}) in Table 5.6, are

[18]The squared variables (l^2 and w_l^2) are introduced in order to capture the potential nonlinearity. However, we cannot plug the productivity measurements into the nonlinear function, because they are estimated variables.

significantly negative for all periods. These estimates correspond to the (negative) sunk cost of entry, which represents the expectation of the difference between the treated group (newly entered firms) and the control group (established exporters) in terms of export revenue. Therefore, the negative and significant estimated coefficients of D_{it-1} suggest the existence of sunk costs of entry for the newly entered firms.[19] The alternative method - the propensity score matching, provides very similar results, see Table 5.9 in Appendix B. We also examine the impact of productivity on the sunk cost. The estimated coefficients of variable $D_{it-1} \cdot TFP_{it-1}$ are significantly positive for all panels, which indicate that the cost of entry is decreasing with TFP. In other words, this suggests that firms with higher TFP levels incur significantly less entry costs than the average. This empirical evidence supports our theoretical framework, in which the sunk cost of entry is modeled as a function of firm-level productivity. The empirical results regarding the impact of RFP are less conclusive. Although the estimated coefficients of variable $D_{it-1} \cdot RFP$ are negative in five out of six cases, the bootstrap confidence intervals show that they are not precisely estimated.

5.5 Conclusion

Estimation results in this chapter suggest that the firms in our sample face a significant sunk cost of entry into the foreign market and there is a significant link between productivity and sunk costs. These results validate the working assumption made in our theoretical model: sunk entry costs are heterogeneous across firms and less important for more productive firms. We show theoretically that the productivity elasticity of entry costs (γ) has a great impact on the minimum entry requirement at equilibrium and therefore on the self-selection process. We find that, in a selective export market, a higher elasticity makes the market entry easier. The opposite occurs in a more open market.

The limitation of the current model is that only the entry of firms is considered while the exit decision is disregarded. Further research may take a direction toward a more dynamic trade model, where productivity evolves over time as a stochastic process (see Costantini and Melitz, 2007). Thus, we can model firm's decision explicitly for each period of time, which will allow us to analyze the exit process.

[19] The estimates of λ_1 are not interpretable in our model, they correspond to the linear prediction of a composite function, i.e., $G_0(x_{it-1}) \equiv f(x_{it-1}) + g_0(x_{it-1})$.

5.6 Appendix

Appendix A: Theoretical model derivations

Appendix A.1: Proof of Proposition 5.1 - Derivation of the Equilibrium

Before presenting derivation of equilibrium, we provide a summary of all parameters used in our theoretical model and the corresponding restrictions.

Table 5.7: Summary of parameters

Notation	Definition	Support
$\bar{\varphi}$	scale parameter of Pareto dist.	$[0, +\infty[$
k	slope parameter of Pareto dist.	$[0, +\infty[$
ϵ	elasticity of substitution across varieties	$[1, +\infty[$
γ	productivity elasticity of entry costs	$[0, +\infty[$
f_X	export fixed costs	$]0, +\infty[$
f_E	domestic entry costs	$]0, +\infty[$
f_D	production fixed costs	$]0, +\infty[$
τ	iceberg costs	$[0, +\infty[$
β	$k/\epsilon - 1$	$[1, +\infty[$
Δ	$\beta - (\beta - 1)[k/(k + \gamma)]$	$[0, +\infty[$
Ω	indicator of openness, $\Omega \equiv (\tau^{1-\epsilon})^{\beta}(f_X/f_D)^{1-\beta}$	$[0, 1]$

To solve for the equilibrium values of φ_D, φ_X and Λ, we use the free entry condition and the zero cutoff profit condition for both domestic and export market:

$$\Lambda \varphi_D^{1-\epsilon} = f_D \tag{5.19}$$

$$\Lambda(\tau\varphi_X)^{1-\epsilon} = f_X h(\varphi) \tag{5.20}$$

$$V(\varphi_D) \cdot \Lambda + \tau^{1-\epsilon} V(\varphi_X) \cdot \Lambda - \left[G(\varphi_D) f_D + f_X \int_0^{\varphi_X} h(\varphi) dG(\varphi) \right] = f_E. \tag{5.21}$$

Using the definition of the expected value of firms, we can solve for $V(\varphi_i)$ $(i = D, X)$:

$$V(\varphi_i) = \int_0^{\varphi_i} \varphi^{1-\epsilon} dG(\varphi) = \int_0^{\varphi_i} \varphi^{1-\epsilon} g(\varphi) d\varphi = \left(\frac{\beta}{\beta - 1} \right) \bar{\varphi}^{-k} \varphi_i^{k-\epsilon+1}. \tag{5.22}$$

From (5.19) and (5.20), we can find the relative equilibrium threshold value for entry:

$$\frac{\varphi_X}{\varphi_D} = \left(\frac{f_X h(\varphi_X)}{f_D \tau^{1-\epsilon}} \right)^{\frac{1}{1-\epsilon}} \tag{5.23}$$

Also, from (5.19) and (5.20) we have:

$$\Lambda = f_D \varphi_D^{\epsilon-1} = f_X h(\varphi_X)(\tau\varphi)^{\epsilon-1} \tag{5.24}$$

Equation (5.21) can be rewritten as:

$$f_E + f_D G(\varphi_D) + f_X \int_0^{\varphi_X} h(\varphi) g(\varphi) d\varphi = B\left[V(\varphi_D) + \tau^{1-\epsilon} V(\varphi_X)\right]$$

Using (5.24) and dividing both sides by $f_D G(\varphi_D)$ yields:

$$\frac{f_E}{f_D G(\varphi_D)} + 1 + \frac{f_X \int_0^{\varphi_X} h(\varphi) g(\varphi) d\varphi}{f_D G(\varphi_D)} = \varphi_D^{\epsilon-1}\left[\frac{V(\varphi_D)}{G(\varphi_D)} + \tau^{1-\epsilon}\frac{V(\varphi_X)}{G(\varphi_D)}\right]$$

Replacing $V(\varphi_D)$ and $V(\varphi_X)$ by (5.22) and using the definition of $G(\varphi) = (\varphi/\bar{\varphi})^k$, we can write:

$$\frac{f_E}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k + 1 + \frac{f_X}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k \int_0^{\varphi_X} h(\varphi) g(\varphi) d\varphi = \left(\frac{\beta}{\beta-1}\right)\left[1 + \tau^{1-\epsilon}\left(\frac{\varphi_X}{\varphi_D}\right)^{k-\epsilon+1}\right]$$

where $\beta \equiv k/\epsilon - 1$. Then using (5.23) we find:

$$\frac{f_E}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k + 1 + \frac{f_X}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k \int_0^{\varphi_X} h(\varphi) g(\varphi) d\varphi = \left(\frac{\beta}{\beta-1}\right)\left[1 + \tau^{1-\epsilon}\left(\frac{f_X h(\varphi_X)}{f_D \tau^{1-\epsilon}}\right)^{\frac{k-\epsilon+1}{1-\epsilon}}\right]$$

$$\frac{f_E}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k + 1 + \frac{f_X}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k \int_0^{\varphi_X} h(\varphi) g(\varphi) d\varphi = \left(\frac{\beta}{\beta-1}\right)\left[1 + \Omega h(\varphi_X)^{1-\beta}\right]$$

where $\Omega \equiv (\tau^{1-\epsilon})^\beta (f_X/f_D)^{1-\beta}$. Now consider the functional form $h(\varphi) = \varphi^\gamma$. We have

$$f_X \int_0^{\varphi_X} h(\varphi) g(\varphi) d\varphi = f_X G(\varphi_X) \frac{k}{k+\gamma} \varphi^\gamma.$$

This yields:

$$\frac{f_E}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k + 1 + \frac{f_X}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k \frac{k}{k+\gamma}(\varphi_X^\gamma) G(\varphi_X) = \left(\frac{\beta}{\beta-1}\right)\left[1 + \Omega(\varphi_X^\gamma)^{1-\beta}\right]$$

$$\frac{f_E}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k + 1 + \frac{f_X}{f_D}\left(\frac{\varphi_X}{\varphi_D}\right)^k \frac{k}{k+\gamma}(\varphi_X^\gamma) = \left(\frac{\beta}{\beta-1}\right)\left[1 + \Omega(\varphi_X^\gamma)^{1-\beta}\right]$$

$$\frac{f_E}{f_D}\left(\frac{\bar{\varphi}}{\varphi_D}\right)^k + 1 + \Omega\frac{k}{k+\gamma}(\varphi_X^\gamma)^{1-\beta} = \left(\frac{\beta}{\beta-1}\right)\left[1 + \Omega(\varphi_X^\gamma)^{1-\beta}\right]$$

$$\frac{f_D}{f_E}\left[\left(\frac{\beta}{\beta-1}\right)\left[1 + \Omega(\varphi_X^\gamma)^{1-\beta}\right] - 1 - \Omega\frac{k}{k+\gamma}(\varphi_X^\gamma)^{1-\beta}\right] = \left(\frac{\bar{\varphi}}{\varphi_D}\right)^k$$

$$\frac{f_D}{f_E(\beta-1)}\left[1 + \Omega(\varphi_X^\gamma)^{1-\beta}(\beta - (\beta-1)\frac{k}{k+\gamma})\right] = \left(\frac{\bar{\varphi}}{\varphi_D}\right)^k$$

Finally we obtain Equation (5.5):

$$\varphi_D^k = \bar{\varphi}^k \frac{f_E(\beta-1)}{f_D[1+\Omega(\varphi_X^\gamma)^{1-\beta}\Delta]},$$

where $\Delta \equiv \beta - (\beta-1)[k/(k+\gamma)]$. Using (5.4) and the specification of $h(\varphi)$, we derive the equilibrium value for φ_X:

$$\begin{aligned}
\frac{\varphi_X^{k+\gamma\beta}\left(\frac{f_X}{f_D\tau^{1-\epsilon}}\right)^\beta}{\bar{\varphi}^k} &= \frac{f_E(\beta-1)}{f_D[1+\Omega(\varphi_X^\gamma)^{1-\beta}\Delta]} \\
\varphi_X^{k+\gamma\beta} &= \bar{\varphi}^k \frac{f_D}{f_X}\frac{f_E(\beta-1)}{f_D[1+\Omega(\varphi_X^\gamma)^{1-\beta}\Delta]}\frac{f_X}{f_D}\left(\frac{f_X}{f_D\tau^{1-\epsilon}}\right)^{-\beta} \\
\varphi_X^{k+\gamma\beta} &= \bar{\varphi}^k \frac{f_E(\beta-1)}{f_X[1+\Omega(\varphi_X^\gamma)^{1-\beta}\Delta]}\Omega \\
\varphi_X^{k+\gamma\beta} + \Omega\Delta\varphi_X^{k+\gamma} &= \bar{\varphi}^k \frac{f_E(\beta-1)\Omega}{f_X} \\
\varphi_X^k &= \bar{\varphi}^k \frac{f_E(\beta-1)\Omega}{f_X(\varphi_X^{\gamma\beta}+\Omega\Delta\varphi_X^\gamma)}
\end{aligned}$$

Note that if $\gamma = 0$ our model corresponds to the linear model because $\Delta = 1$ if $\gamma = 0$.

$$\varphi_D = \bar{\varphi}\left[\frac{f_E(\beta-1)}{f_D(1+\Omega)}\right]^{\frac{1}{k}};\ \varphi_X = \bar{\varphi}\left[\frac{f_E(\beta-1)\Omega}{f_X(1+\Omega)}\right]^{\frac{1}{k}}.$$

Appendix A.2: Proof of Proposition 5.1 - Uniqueness of the Equilibrium

To show the uniqueness, we examine the monotonicity of the implicit function (5.6) that characterizes the equilibrium of φ_X:

$$\bar{\varphi}^k \frac{f_E(\beta-1)\Omega}{f_X(\varphi_X^{\gamma\beta}+\Omega\Delta\varphi_X^\gamma)} - \varphi_X^k = 0,$$

The derivative w.r.t. φ_X is:

$$-\frac{\bar{\varphi}^k f_E(\beta-1)\Omega}{f_X}\frac{(\gamma\beta\varphi_X^{\gamma\beta-1}+\Omega\Delta\gamma\varphi_X^{\gamma-1})}{(\varphi_X^{\gamma\beta}+\Omega\Delta\varphi_X^\gamma)^2} - k\varphi_X^{k-1} < 0.$$

The implicit function (5.6) is monotone. Thus, there is a unique solution for φ_X. Since the equilibrium value of other variables is defined by φ_X, we can conclude that the equilibrium is unique.

Appendix A.3: Proof of Proposition 5.2

φ_X is implicitly defined as a solution to:

$$F(\varphi_X,\gamma)=\bar{\varphi}^k\frac{f_E(\beta-1)\Omega}{f_X(\varphi_X^{\gamma\beta}+\Omega\Delta\varphi_X^{\gamma})}-\varphi_X^k=0,$$

Using the Implicit Function Theorem, we obtain:

$$\frac{\partial\varphi_X}{\partial\gamma}=-\frac{\partial F/\partial\gamma(\varphi_X,\gamma)}{\partial F/\partial\varphi_X(\varphi_X,\gamma)},$$

where

$$\begin{aligned}\frac{\partial F}{\partial\gamma}(\varphi_X,\gamma) &= -\frac{\bar{\varphi}^k f_E(\beta-1)\Omega}{f_X}\frac{1}{(\varphi_X^{\gamma\beta}+\Omega\Delta\varphi_X^{\gamma})^2}\\ &\quad\left[\beta\log(\varphi_X)\varphi_X^{\gamma\beta}+\Omega\frac{\varphi_X^{\gamma}[\log(\varphi_X)(k+\gamma)(\beta\gamma+k)+(\beta-1)k]}{(k+\gamma)^2}\right]\\ &= \Theta\cdot\left[\beta\log(\varphi_X)\varphi_X^{\gamma\beta}+\Omega\frac{\varphi_X^{\gamma}[\log(\varphi_X)(k+\gamma)(\beta\gamma+k)+(\beta-1)k]}{(k+\gamma)^2}\right];\end{aligned}$$

$$\begin{aligned}\frac{\partial F}{\partial\varphi_X}(\varphi_X,\gamma) &= -\frac{\bar{\varphi}^k f_E(\beta-1)\Omega}{f_X}\frac{(\gamma\beta\varphi_X^{\gamma\beta-1}+\Omega\Delta\gamma\varphi_X^{\gamma-1})}{(\varphi_X^{\gamma\beta}+\Omega\Delta\varphi_X^{\gamma})^2}-k\varphi_X^{k-1}\\ &= \Theta\cdot(\gamma\beta\varphi_X^{\gamma\beta-1}+\Omega\Delta\gamma\varphi_X^{\gamma-1})-k\varphi_X^{k-1},\end{aligned}$$

where

$$\Theta\equiv-\frac{\bar{\varphi}^k f_E(\beta-1)\Omega}{f_X}\frac{1}{(\varphi_X^{\gamma\beta}+\Omega\Delta\varphi_X^{\gamma})^2}$$

In order to conclude on the sign of the slope, $\partial\varphi_X/\partial\gamma$, we have to discuss the sign of $\partial F/\partial\gamma$ and $\partial F/\partial\varphi_X$. We start with the latter. First, Θ is negative given a reasonable values of parameters. Second, we note that $\gamma\beta\varphi_X^{\gamma\beta-1}+\Omega\Delta\gamma\varphi_X^{\gamma-1}$ and $k\varphi_X^{k-1}$ are positive. Thus, we conclude that $\partial F/\partial\varphi_X$ is negative.

Sign of $\partial F/\partial\gamma$: The sign of $\partial F/\partial\gamma$ is less obvious. First, if we only consider the interval $\varphi_X\in[1,+\infty[$, $\partial F/\partial\gamma$ is negative. In contrast, for the interval $\varphi_X\in[0,1[$, $\log(\varphi_X)$ is negative. Thus, the sign of$\partial F/\partial\gamma$ is indeterminate. However, we can conclude on the sign of $\partial\varphi_X/\partial\gamma$ or the sign of

$$\beta\log(\varphi_X)\varphi_X^{\gamma\beta}+\Omega\frac{\varphi_X^{\gamma}[\log(\varphi_X)(k+\gamma)(\beta\gamma+k)+(\beta-1)k]}{(k+\gamma)^2},$$

if we can find a condition on φ_X that ensure $\log(\varphi_X)(k+\gamma)(\beta\gamma+k)+(\beta-1)k$ to be negative. This is because $\beta\log(\varphi_X)\varphi_X^{\gamma\beta}$ is always negative for $\varphi_X\in[0,1[$, see Table (5.7). This condition is:

$$\log(\varphi_X)<-\frac{(\beta-1)k}{(k+\gamma)(\beta\gamma+k)}=\frac{1}{k+\gamma}-\frac{1}{\epsilon-1+\gamma}<1.$$

Therefore we have:

- i) For $\varphi_X \in [1, +\infty[$, φ_X is decreasing in γ, i.e., the slope $\partial\varphi_X/\partial\gamma > 0$;
- ii) For $\varphi_X \in]0, \exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma})]$, φ_X is increasing in γ, i.e., the slope $\partial\varphi_X/\partial\gamma < 0$;
- iii) For $\varphi_X \in]\exp(\frac{1}{k+\gamma} - \frac{1}{\epsilon-1+\gamma}), 1[$, the sign of slope is undetermined.

Appendix B: Estimation results of productivity and Robustness checks

Table 5.8: Correlation matrices

2004-5	$\widehat{TFP}_{t-1}^{CB}$	$\widehat{TFP}_{t-1}^{CES}$	$\widehat{RFP}_{t-1}^{CES}$	Y_{t-1}	L_{t-1}	K_{t-1}
$\widehat{TFP}_{t-1}^{CES}$	0.935					
$\widehat{RFP}_{t-1}^{CES}$	0.057	0.327				
Y_{t-1}	0.844	0.734	0.166			
L_{t-1}	0.383	0.173	0.046	0.722		
K_{t-1}	0.323	0.364	0.724	0.623	0.719	
r_t	0.048	-0.013	0.170	0.306	0.504	0.484

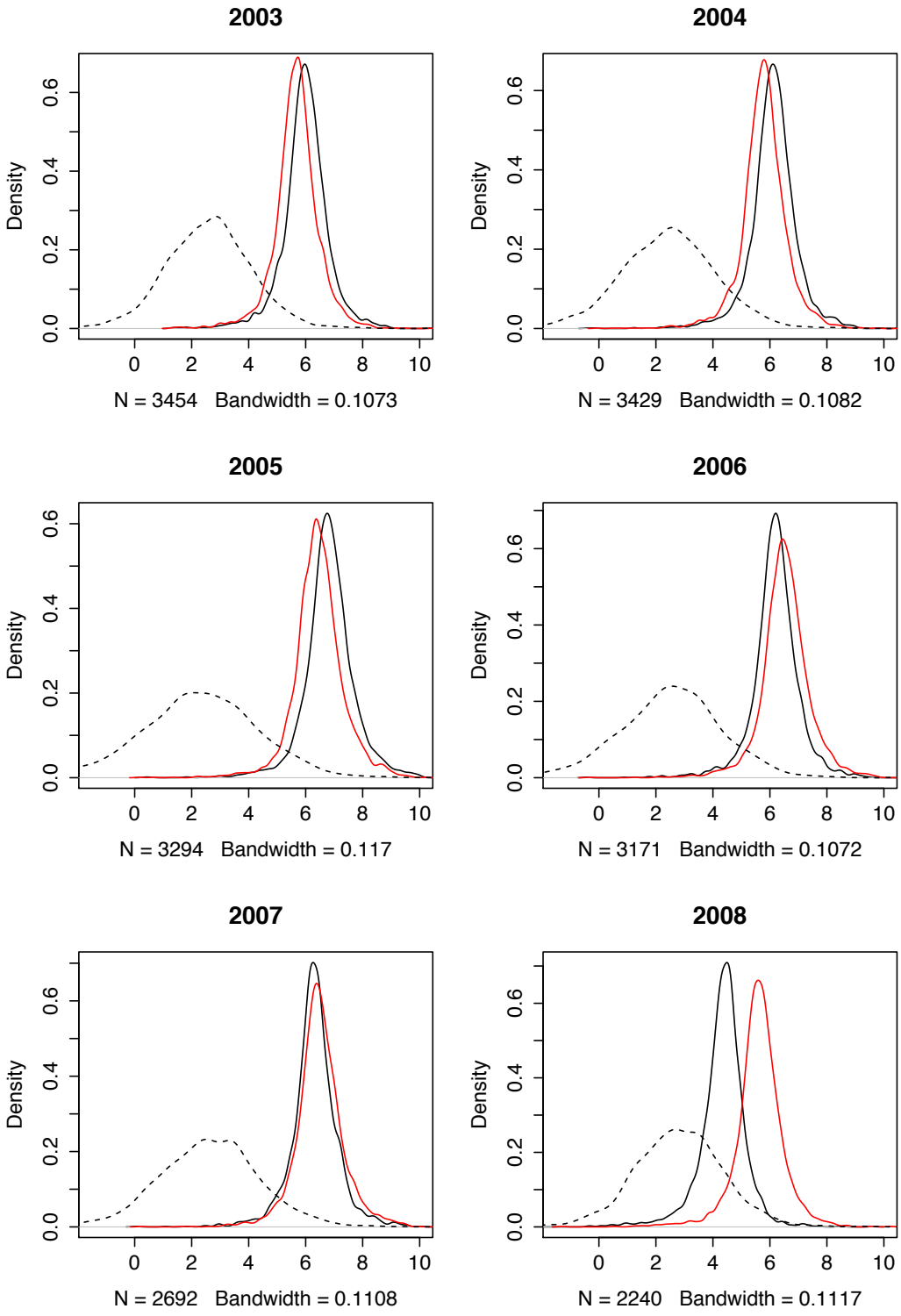
2005-6	$\widehat{TFP}_{t-1}^{CB}$	$\widehat{TFP}_{t-1}^{CES}$	$\widehat{RFP}_{t-1}^{CES}$	Y_{t-1}	L_{t-1}	K_{t-1}
$\widehat{TFP}_{t-1}^{CES}$	0.947					
$\widehat{RFP}_{t-1}^{CES}$	0.150	0.382				
Y_{t-1}	0.893	0.801	0.175			
L_{t-1}	0.460	0.277	0.037	0.708		
K_{t-1}	0.446	0.481	0.726	0.622	0.708	
r_t	0.113	0.048	0.137	0.300	0.516	0.475

2006-7	$\widehat{TFP}_{t-1}^{CB}$	$\widehat{TFP}_{t-1}^{CES}$	$\widehat{RFP}_{t-1}^{CES}$	Y_{t-1}	L_{t-1}	K_{t-1}
$\widehat{TFP}_{t-1}^{CES}$	0.960					
$\widehat{RFP}_{t-1}^{CES}$	0.056	0.318				
Y_{t-1}	0.849	0.838	0.189			
L_{t-1}	0.366	0.327	0.060	0.706		
K_{t-1}	0.308	0.462	0.736	0.620	0.715	
r_t	0.029	0.053	0.156	0.281	0.487	0.458

2007-8	$\widehat{TFP}_{t-1}^{CB}$	$\widehat{TFP}_{t-1}^{CES}$	$\widehat{RFP}_{t-1}^{CES}$	Y_{t-1}	L_{t-1}	K_{t-1}
$\widehat{TFP}_{t-1}^{CES}$	0.962					
$\widehat{RFP}_{t-1}^{CES}$	0.083	0.331				
Y_{t-1}	0.857	0.831	0.196			
L_{t-1}	0.402	0.331	0.060	0.715		
K_{t-1}	0.352	0.474	0.732	0.634	0.720	
r_t	0.067	0.069	0.165	0.303	0.506	0.480

2008-9	$\widehat{TFP}_{t-1}^{CB}$	$\widehat{TFP}_{t-1}^{CES}$	$\widehat{RFP}_{t-1}^{CES}$	Y_{t-1}	L_{t-1}	K_{t-1}
$\widehat{TFP}_{t-1}^{CES}$	0.844					
$\widehat{RFP}_{t-1}^{CES}$	-0.260	0.293				
Y_{t-1}	0.608	0.725	0.187			
L_{t-1}	0.000	0.053	0.038	0.650		
K_{t-1}	-0.170	0.264	0.746	0.577	0.691	
r_t	-0.164	-0.054	0.153	0.268	0.474	0.445

Figure 5.4: Distributions of estimated productivity



Note: the black line is the distribution of TFP based on Cobb-Douglas; the red line is the distribution of TFP based on CES; the dash line is the distribution of RFP.

Table 5.9: Estimates of ATE by Propensity score matching

	2003-04	2004-05	2005-06	2006-07	2007-08	2008-09
Intercept	−5.710 (0.896)	−4.995 (1.055)	−1.500 (1.053)	−1.940 (1.035)	−5.085 (1.126)	−3.207 (1.219)
Sector	$2.867e^{-4}$ ($1.187e^{-4}$)	$3.367e^{-4}$ ($1.303e^{-4}$)	$1.262e^{-4}$ ($1.365e^{-4}$)	$1.048e^{-5}$ ($1.273e^{-4}$)	$5.953e^{-5}$ ($1.480e^{-4}$)	$1.686e^{-4}$ ($1.563e^{-4}$)
TFP_{t-1}	0.809 (0.199)	0.882 (0.222)	0.699 (0.224)	1.152 (0.2339)	0.714 (0.256)	1.043 (0.288)
RFP_{t-1}	1.739 (0.592)	1.979 (0.514)	1.338 (0.289)	1.939 (0.401)	1.684 (0.500)	3.285 (0.815)
Labor_{t-1}	2.235 (0.843)	3.036 (0.828)	2.050 (0.557)	3.146 (0.673)	2.236 (0.809)	4.657 (1.179)
Capital_{t-1}	−2.670 (0.848)	−3.390 (0.826)	−2.753 (0.569)	−3.428 (0.678)	−2.832 (0.819)	−4.816 (1.165)
Output_{t-1}	0.301 (0.180)	0.276 (0.201)	0.196 (0.212)	−0.193 (0.219)	0.425 (0.232)	0.033 (0.248)
ATE	−1.603 (0.408)	−2.484 (0.457)	−0.501 (0.498)	−2.165 (0.64383)	−1.680 (0.580)	−1.125 (0.617)

Note: the propensity score is computed using a logit regression in the first-stage. The logistic regression results are reported in Lines 2-8.

Chapter 6

General Conclusion

The aim of this thesis has been to identify and extend some limits of recent contributions to production behavior modeling. In the preceding chapters, I compared different empirical specifications and statistical methods which have often been used in production analysis, and pointed out their implications for estimating technology parameters. I have studied the causes and cures of the endogeneity problem in the context of production analysis. This thesis also addressed the important but neglected issue of fixed costs. This work defined and characterized the fixed cost, and developed empirical strategies to estimate the fixed cost using the standard production database. Empirical evidence suggests that the fixed cost is significant and has profound impacts on producer's behavior in terms of price setting, returns to scale and exports.

6.1 Main findings and implications of the analysis

This thesis contributes to the theory and empirics of production analysis on the following five points:

- Estimating returns to scale, elasticity of substitution and technical change;
- Problem of endogeneity;
- Productivity measurements;
- Microeconomic foundation of fixed costs;
- Identification of fixed costs.

Estimating technology parameters is a primary objective of applied production analysis. However, divergent and often contradictory estimation results are documented in the literature which mislead and hamper the development of the field. Chapter 2, 3 and 4 of this thesis identify a series of reasons for the lack of consensus on three key parameters: returns to scale, elasticity of substitution and technical change. First, I

show in Chapter 2 that the direct and reverse regressions yield two different estimation results on the degree of returns to scale, and this difference is due to the assumptions on the stochastic specification of error terms. A second reason for the divergence of estimates is highlighted in Chapter 3. Often, the estimates of parameters based on different economic assumptions are compared. In many empirical works, the estimates are obtained by imposing restrictions on key parameters, such as constant returns to scale, unitary elasticity of substitution or Hicks-neutral technical change. The estimation method proposed in Chapter 3 avoids this source of bias by estimating simultaneously the three parameters in a semi-parametric setup. Third, we also point out in Chapter 4 that the parameters such as returns to scale may be overestimated when the fixed cost is neglected.

Beside the estimation bias that is introduced by stochastic assumptions and functional restrictions, the **problem of endogeneity** is another source of bias in empirical production analysis. In Chapter 2, both input and output growth variables in the model of interest are measured with error. Thus, endogeneity arises as a result of measurement error. The IV estimation is a solution to this problem. However, I show that the use of weak instruments may, in contrary, magnify this bias. Therefore, a series of tests for weak instruments are needed and our study suggests that alternative IV estimations as the LIML and F-LIML estimators are more robust against weak instruments. In a different context (Chapter 3), estimation of production function, the endogeneity problem is caused by a loop of causality between the input demand and output variables through the unobserved productivity terms. The classic IV estimation can be used to deal with this problem, but the valid instruments are not always available in this context. Based on the Cobb-Douglas production function, an alternative and more structural method, the so-called control function approach, is introduced by Olley and Pakes (1996). In Chapter 3, I extended this approach to a CES production function, allowing for complementarity between production factors and biased technical change.

The study on the two preceding points also provides a solid base for investigating **productivity measurements**. The measures of productivity (Hicks-neutral as well as non-neutral) are obtained as the residuals in the functional relationships that reflect both firms' production technology and optimization behavior. Thus, consistent estimates of parameters in these functions are essential for recovering productivity. Traditional productivity studies only focus on the measure of TFP (Hicks-neutral) obtained as the residual of a Cobb-Douglas production function. Based on a CES production function, the method proposed in Chapter 3 yields an individual-specific measure of RFP (Relative Factor-augmenting Productivity) by incorporating the additional information on firms' optimization behavior into the estimation. This new measure of non-neutral productivity opens the possibility for further understanding of technical change and production efficiency, and can be used for various applications. For example, we use

this extension in Chapter 5 to investigate the relationship between costs of entry into export markets and productivity.

Chapter 4 and 5 discuss the **microeconomic foundation of fixed costs**. We extend in Chapter 4 the classical production function in order to generate a fixed cost. This fixed cost is defined as the cost of inputs required for producing an arbitrarily small amount of output. We study how fixed costs affect firms' behavior in terms of price setting and returns to scale. In Chapter 5, we consider another type of fixed costs: the entry cost into the export market. We propose a theoretical trade model that highlights the relationship between productivity and entry costs. Our model shows that this relationship affects the minimum entry requirement at equilibrium and has an impact on the self-selection process. After having provided definitions and studied implications of fixed costs, our analysis moves toward the empirical question: how to identify fixed costs from data?

The **identification of fixed costs** conformably to their specific definition, is challenging by the fact that most of standard production data do not provide information on the fixed input and fixed cost. Chapter 4 and 5 contribute to this point by introducing two identification strategies. In Chapter 4, we propose an empirical cost function that is two-point flexible, has a fixed and a variable component and allows cost heterogeneity. Empirical results based on this cost function suggest that a considerable part of production costs is fixed, and there is a trade-off between the fixed and variable cost. We also find that industries with higher fixed costs, produce more in average, benefit from a higher markup as well as returns to scale, and are more concentrated. In Chapter 5, we estimate the cost of entry into the export market, which is quantified in terms of export revenue shortfall. Our empirical strategy consists of comparing the revenue of newly entered exporters with those of established exporters, based on a treatment evaluation model. We find evidence that entry costs exist and more productive firms are able to enter the market with less entry costs.

6.2 Limitations and extensions

This work contributes to the literature in several aspects as discussed in the previous section. However, it also encountered a number of limitations, which need to be considered. These issues can be grouped into four categories, discussed below.

Model specification. In Chapter 2, I follow the Diewert and Fox's (2008) method of measuring technical progress and returns to scale. As a direct consequence of this methodology, the output growth index is linearly related to the input growth index. An extension of this study is to generalize the analysis to nonlinear specifications of the production technology along the lines of Kumbhakar and Tsionas (2011). In Chapter 3, I extend Olley and Pakes's (1996) method to a CES production function by allowing

non-unitary substitution elasticity between factors. However, the basic CES form still restricts the substitution elasticity to be constant and it is not suitable for more than two inputs in the production technology, see Uzawa (1962) and Sato (1967) on this point. Therefore, the extension of this approach to flexible functional forms is the topic of both theoretical and practical interest for future research. Note that the ongoing work on nonparametric identification and estimation of production functions using the control function approach offers further research possibilities (see Hong, 2008).

Level of aggregation. The analysis in Chapter 2 and 3 are conducted at different levels of aggregation: two-digit SIC for Chapter 2 and six-digit NAICS for Chapter 3. A number of studies show that the level of aggregation may affect the estimation results of returns to scale and of substitution elasticities (Solow, 1964, p.118 and Basu, 2008). Therefore, it is interesting to evaluate the effects of aggregation on the estimates of technology parameters. We can for example use the NBER database (six-digit NAICS) to reconstruct series for different levels of aggregation and compare the corresponding estimation results. The increasing availability of firm-level data offers another possible extension to this thesis. In Chapter 4, we estimate the fixed cost at the sector-level. The same analysis can be carried out in the future at the firm-level. Such a study would allow us to examine explicitly strategic interactions between firms in their joint decision on product price and production capacity. This would expand our understanding of the link between fixed cost and barriers to entry.

Stochastic properties of productivity and unified trade model. The idea that firms improve their productivity through exporting is referred to as learning-by-exporting. Numerous studies have been devoted to test empirically this hypothesis. However, no clear evidence has been found to support learning-by-exporting. A recent paper by De Loecker (2013) provides some answers to this question. De Loecker (2013) argues that the current econometric methods (including those used in this thesis) rely on the assumption that productivity evolves exogenously, which is not suitable for characterizing learning-by-exporting. He introduces a new estimation method that incorporates endogenous productivity processes. A possible extension of this thesis is to generalize De Loecker's (2013) proposition to our method in Chapter 3. This would provide the empirical basis for a key element of our research agenda: *unified trade theory*. In Chapter 5, our theoretical model focused only on the self-selection process, where the effect of learning-by-exporting is not modeled. In the spirit of Costantini and Melitz (2007), we could extend our trade model to a more dynamic one, which incorporates explicitly entry, exit, technology choices and endogenous productivity in every period. This model would unify the two features of the export premium: self-selection and learning-by-exporting, identify the different channels of aggregate productivity improvements, and provide recommendations for outward-oriented policy reforms.

Chapitre 7

Résumé de thèse

(Summary of the thesis in French)

Cette thèse se compose de quatre chapitres sur l'analyse de la production appliquée, avec un accent sur la technologie, la productivité et les coûts fixes.

7.1 Contexte

Dans *Value and Capital* (1946), Hicks présente la théorie de la production comme l'étude des relations entre les facteurs de production (input) et les biens produits (output), ainsi que celle des relations entre les facteurs de production. Afin de caractériser ces relations, la théorie est basée sur l'hypothèse que les entreprises minimisent le coût de production et maximisent le profit, sous la contrainte technologique.

Le premier objectif de l'analyse de la production est de formaliser le comportement des producteurs à l'aide de modèles mathématiques, et de les représenter par les notions de substituabilité des inputs, de rendement d'échelle, et de progrès technique. L'objectif second de cette analyse est de mesurer empiriquement ces concepts en utilisant les données de production et des méthodes statistiques.

Historiquement, l'analyse appliquée de la production est née du travail de Cobb et Douglas (1928) intitulé "*A Theory of Production*". Dans leur approche, la technologie de production est représentée par une fonction paramétrique. Après un travail minutieux de la collecte des données, Cobb et Douglas ont estimé leur modèle sur données manufacturières américaines pour la période de 1899 à 1922. Ils "matérialisent", pour la première fois, la relation input-output :

$$P' = 1.01 L^{3/4} C^{1/4}.$$

D'après les notations utilisées par Cobb-Douglas (1928), P' désigne la valeur ajustée de la production, L et C désignent le travail et le capital, respectivement. Par rapport aux techniques d'aujourd'hui, la méthode statistique utilisée par Cobb et Douglas (1928) semble restrictive. Dans les décennies suivantes, une série de contributions à l'analyse de la production ont été réalisés. Frisch (1935) a tenté de mesurer les possibilités de substitution entre les facteurs de production dans la fabrication du chocolat. Arrow, Chenery, Minhas et Solow (1961) ont proposé la fonction de production, communément appelée Constant Elasticity of Substitution (CES) *production function*. Hotelling (1932), Samuelson (1960) et Shephard (1970) ont introduit la théorie de la dualité. Diewert (1971), Christensen, Jorgenson et Lau (1973) ont proposé des formes fonctionnelles flexibles pour la modélisation de la technologie de production.

Accompagnée par la démocratisation des données et de l'augmentation de la puissance de calcul au cours du siècle dernier, l'analyse empirique de la production a été considérablement améliorée. Malgré d'importants progrès dans ce domaine, nous sommes toujours confrontés à de nombreux défis théoriques et empiriques (cf. Jorgenson, 1986 ; Griliches et Mairesse, 1995 et Ackerberg et al., 2007). L'objectif principal de cette thèse est de développer des stratégies empiriques en mettant l'accent sur les spécifications empiriques de la technologie de production et les méthodes d'estimation.

7.2 Problématiques

Cette thèse vise à étudier les techniques récemment mises au point pour estimer les fonctions de coûts et de production afin de souligner leurs limites, et d'apporter des améliorations à ces méthodes. Un grand nombre d'articles décrit les questions d'estimation de l'approche primitive (*the primal approach*), mais une solution alternative, connue sous le terme d'approche duale (*the duality approach*), est également utilisée. La déduction des fonctions de demande à partir de la fonction de production est bien connue, selon la théorie de la dualité. Il est également possible de déduire la fonction de production par le système des fonctions de demande, car ces deux notions incorporent les mêmes informations sur les technologies. Alors qu'une telle relation de dualité est théoriquement bien connue, l'économétrie de la dualité n'a pas vraiment été étudiée. Il est intéressant par exemple d'étudier les conditions dans lesquelles l'estimation des relations technologiques, par l'approche duale, est plus adaptée que l'approche primitive (et inversement). Les chapitres 2 à 5 de la thèse contribuent à l'extension de la littérature concernant l'analyse de la production existant en traitant plus spécifiquement les questions suivantes :

Spécification stochastique Bien que la forme et les propriétés des fonctions de production et de coût peuvent être déduites de la théorie, la théorie reste muette sur la manière dont les termes des erreurs doivent être inclus dans le modèle. Quelles sont les implications de différentes spécifications stochastiques ? Voir le chapitre 2.

Substitution des facteurs de production Malgré la simplicité de la fonction CES, les différents modèles empiriques ont produit des résultats divergents sur la valeur de l'élasticité de substitution. Dans quelle mesure ces nouvelles techniques d'estimation permettent d'apporter de nouveaux éclairages sur cette élasticité ? Voir le chapitre 3.

Biais de progrès techniques Pour la majorité des spécifications qui sont utilisées dans les études empiriques, le progrès technique est supposé être neutre au sens de Hicks. Est-il possible d'identifier le biais de progrès techniques ? Et, comment estimer la productivité du facteur augmentant (*factor-augmenting productivity*) ? Voir le chapitre 3.

Coûts fixes et coûts variables La fonction Translog est considérée comme une forme fonctionnelle flexible et utilisée dans de nombreuses études empiriques. Cependant, une hypothèse sous-jacente à cette classe de formes fonctionnelles, est que les inputs sont ajustés instantanément à leur niveau optimal. Est-ce que la fonction de coût Translog est suffisante pour modéliser le coût fixe ? Sinon, quelles en sont les alternatives ? Voir le chapitre 4.

Application dans d'autres domaines Les résultats de l'analyse de la production ont de profondes implications dans d'autres domaines de l'économie, en particulier pour les modèles de commerce international. Dans ces modèles, la productivité et les coûts fixes jouent un rôle crucial pour caractériser l'équilibre, voir Melitz (2003). De nombreux travaux empiriques ont étudié indépendamment les impacts de deux éléments sur le comportement de l'entreprise. Quelle est la relation entre la productivité et les coûts fixes d'entrée ? Est-ce que les entreprises plus productives peuvent entrer sur le marché international avec un frais d'entrée moindre ? Voir le chapitre 5.

7.3 Résumé de chapitres

Chaque chapitre de cette thèse se compose de trois éléments clés de l'analyses économique : la théorie, la formulation empirique et la méthode statistique. Les chapitres 2 à 4 revisitent ces éléments clés et contribuent à un ou plusieurs d'entre eux. le chapitre 5 synthétise et applique les résultats pour les modèles du commerce international. Dans le chapitre de conclusion, les limites de mes travaux et leurs extensions potentielles sont discutés.

La figure suivante fournit un résumé des différents thèmes étudiés dans la thèse. Les thèmes spécifiques sont organisés suivant trois piliers : théorie, modèle mathématique et méthode statistique.

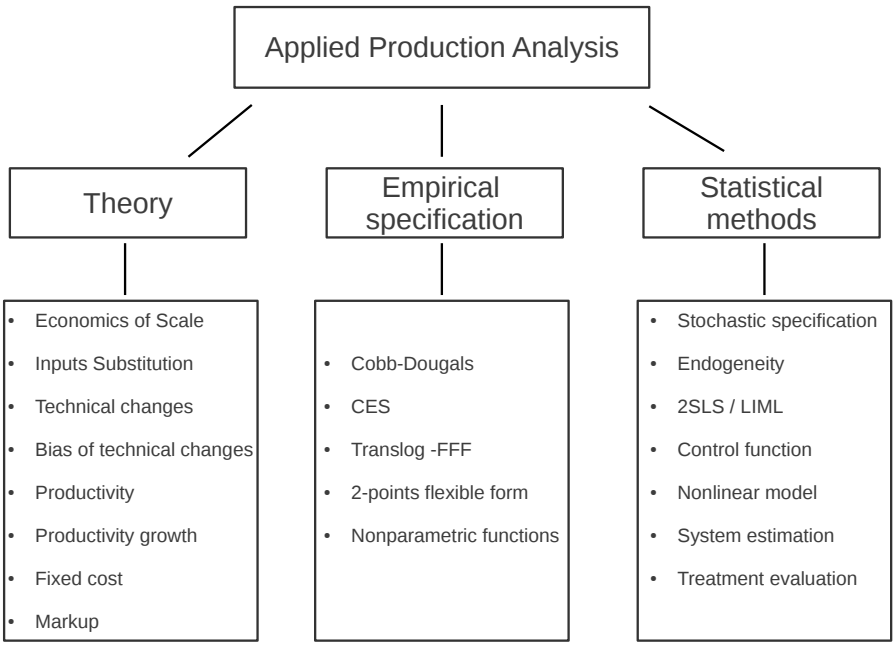


Figure 7.1: Sujets étudiés dans cette thèse

Un effort a été fait dans cette thèse pour unifier (autant que possible) la théorie à la formulation mathématique et la méthode empirique. Il reflète la volonté d'appréhender l'économétrie à la manière d'Ragnar Frisch (1933) : [...] econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory [...] Nor should econometrics be taken as synonymous with the application of mathematics to economics. [...] It is the unification of all three that is powerful. And it is this unification that constitutes econometrics".

Chapitre 2 : "Direct and Reverse Estimates of Returns to Scale"

Dans ce chapitre, nous avons mené une étude comparative de quatre estimateurs : MCO, 2SLS, LIML et F-LIML. Les différents estimateurs sont comparés pour la régression directe et inverse dans le contexte de l'estimation du rendement d'échelle, et du progrès technique. Les résultats montrent que la méthode d'estimation 2SLS (direct/inverse) peut conduire à des conclusions contradictoires lorsque les instruments sont faibles. D'autre part, les estimateurs LIML et F-LIML fournissent des estimations plus fiables. Les résultats basés sur des données de l'industrie manufacturière aux États-Unis, montrent que la plupart des industries présentent des rendements d'échelle croissants et que le rôle du progrès technique est relativement faible.

Chapitre 3 : "Hicks-neutral and Non-neutral Productivity"

Ce chapitre aborde une question importante de l'analyse de la production, à savoir l'estimation de la productivité. Nous proposons une extension de la méthode Olley-Pakes (1996) adaptée à la fonction de production CES avec le progrès technique biaisé. La nouvelle approche semi-paramétrique permet d'estimer le degré de rendements d'échelle, l'élasticité de substitution et le biais de progrès techniques. La méthode proposée a l'avantage non seulement d'estimer la *TFP* (*Total Factor Productivity*), mais également de donner des estimations de la productivité dite factor-augmenting. L'étude empirique montre que les industries manufacturières sont caractérisées par une élasticité de substitution inférieure à l'unité et un biais important de progrès technique.

Chapitre 4 : "Fixed and Variable Cost"

La distinction entre coûts fixes et variables est fondamentale, mais assez négligée dans l'analyse de la production. Peu de papiers ont étudié la substituabilité entre les inputs fixes et variables. Empiriquement, la plupart des spécifications de la fonction de coûts supposent que les coûts fixes sont nuls. Nous proposons donc dans ce chapitre une structure de la fonction de production qui est nécessaire et suffisante pour obtenir un coût fixe. En utilisant les données de l'industrie, nous estimons les éléments fixes et variables de la fonction de coût, et étudions la façon dont les coûts fixes et variables interagissent, ainsi que l'impact sur le comportement des entreprises en matière de *pricing* et des rendements d'échelle.

Chapitre 5 : "Productivity, Fixed Cost and Export"

Le but de ce chapitre est d'étendre le modèle de Melitz (2003) en endogénéisant les coûts d'entrée sur les marchés d'exportation. Notre modèle théorique du commerce international met en évidence le rôle de l'hétérogénéité dans la structure de coûts fixes et les liens entre ces coûts et productivité. Nous montrons que le modèle de Melitz avec des coûts d'entrée exogènes et homogènes ne permet pas d'estimer correctement les mécanismes d'auto-sélection. En réponse à ces résultats théoriques, nous développons une stratégie empirique basée sur les techniques de treatment evaluation pour mesurer les barrières à l'entrée sur le marché d'exportation (avec des coûts endogénéisés). En utilisant des données pour les entreprises françaises, nos investigations empiriques mettent en lumière l'importance des barrières à l'entrée et les déterminants de ces barrières.

7.4 Principaux résultats et les implications de l'étude

Dans cette thèse, nous avons comparé les principales spécifications empiriques et les principales méthodes statistiques qui ont été utilisées dans l'analyse de la production, et souligné leurs implications pour l'estimation des paramètres technologiques. Nous avons étudié les causes et les remèdes du problème d'endogénéité dans le cadre de l'estimation de la fonction de production. Cette thèse a également abordé une question importante mais largement négligée de la littérature : les coûts fixes. Ce travail contribue à la définition et à la caractérisation des coûts fixes. Nous avons développé des stratégies d'estimation du coût fixe, et montré que le coût fixe est important et a un impact significatif sur la politique des prix, les rendements d'échelle et les exportations.

Cette thèse contribue à l'analyse de la production sur les cinq points suivants :

1. Estimation des rendements d'échelle, de l'élasticité de substitution et du changement technique ;
2. Problème d'endogénéité ;
3. Mesures de la productivité ;
4. Fondement microéconomique des coûts fixes ;
5. Identification des coûts fixes.

Estimation des paramètres technologiques L'estimation des paramètres technologiques est un objectif primordial de l'analyse appliqué de la production. Cependant, les

résultats d'estimation que l'on trouve dans la littérature sont divergents, voire contradictoires. Les chapitres 2, 3 et 4 de cette thèse identifient les raisons des divergences de résultats relatifs à trois paramètres clés : les rendements d'échelle, l'élasticité de substitution et le changement technique. Tout d'abord, nous montrons dans le chapitre 2 que les régressions directes et inverses donnent deux résultats d'estimation différentes sur le degré de rendements d'échelle. Cette différence est due aux hypothèses sur la spécification stochastique des termes d'erreur. Une deuxième raison de la divergence des estimations est mis en évidence dans le chapitre 3. Dans de nombreux travaux empiriques, l'estimation d'un paramètre est obtenu en imposant des restrictions, tels que les rendements d'échelle constants, l'élasticité unitaire de substitution ou changement technique neutre. Une méthode d'estimation semi-paramétrique est proposée le chapitre 3, qui permet d'éviter ce biais en estimant simultanément les trois paramètres. Dans le chapitre 4, nous soulignons également que les paramètres tels que les rendements d'échelle peuvent être surestimés lorsque le coût fixe est négligée.

Problème de l'endogénéité En plus des biais d'estimation introduits par des hypothèses stochastiques et des restrictions fonctionnelles, le problème de l'endogénéité est une autre source de biais dans l'analyse de la production. Dans le chapitre 2, l'endogénéité apparaît à la suite d'une erreur de mesure. L'estimation IV est une solution à ce problème. Cependant, nous montrons que l'utilisation d'instruments faibles peut amplifier le biais d'estimation. Par conséquent, une série de tests pour identifier les instruments faibles est recommandée et notre étude suggère que des méthodes alternatives comme le LIML et F-LIML sont plus robustes lorsque les instruments sont faibles. Dans un contexte différent d'estimation de la fonction de production, le problème d'endogénéité est causé par la corrélation entre les termes de productivité non observés et les *inputs*. La méthode d'estimation IV peut aussi être utilisé pour faire face à ce problème, mais les instruments valides sont rarement disponibles dans ce contexte. Une méthode alternative et plus structurelle, est proposée par Olley et Pakes (1996). Dans le chapitre 3, cette approche est étendue à une fonction de production CES, qui permet de prendre en compte la complémentarité entre les facteurs de production et le progrès technique biaisé.

Mesures de la productivité L'analyse des deux points précédents fournit également une base solide pour étudier les mesures de la productivité. Les mesures de la productivité (Hicks-neutre et non-neutre) sont obtenus comme les résidus des fonctions qui reflètent à la fois la technologie de production et le comportement d'optimisation des entreprises. Ainsi, les estimations convergentes des paramètres de ces fonctions sont

essentielles pour mesurer la productivité. Les études traditionnelles de productivité se concentrent uniquement sur la mesure de la *TFP* (Hicks-neutre) obtenu comme le résidu d'une fonction de production. La méthode proposée dans le chapitre 3 donne une mesure de la *RFP* (*Relative Factor-augmenting Productivity*) en intégrant les informations complémentaires sur le comportement d'optimisation des entreprises dans l'estimation. Cette nouvelle mesure de la productivité non-neutre ouvre la possibilité d'études plus approfondie sur le changement technique et l'efficacité de la production, et peut être utilisée pour diverses applications. Par exemple, nous utilisons cette extension dans le chapitre 5 pour étudier la relation entre les coûts d'entrée sur les marchés d'exportation et la productivité.

Fondement microéconomique des coûts fixes Chapitre 4 et 5 traitent du fondement microéconomique des coûts fixes. Dans le chapitre 4, nous étendons la fonction de production classique afin de générer un coût fixe. Ce coût fixe est défini comme le coût des inputs nécessaires à la production d'une arbitrairement petite quantité d'output. Nous étudions comment les coûts fixes influencent le comportement des entreprises en matière de fixation des prix et des rendements d'échelle. Dans le chapitre 5, nous considérons un autre type de coûts fixes : le coût d'entrée sur le marché d'exportation. Nous proposons un modèle théorique qui met en évidence la relation entre les coûts d'entrée et la productivité. Notre modèle montre que cette relation influence la condition d'entrée minimum à l'équilibre et a un impact sur le processus d'auto-sélection du marché. Après avoir donné des définitions et étudié les implications de coûts fixes, notre analyse se déplace vers une question empirique : comment identifier les coûts fixes à partir des données ?

Identification des coûts fixes L'identification des coûts fixes est difficile, car la plupart des données standard ne fournissent pas d'informations sur les facteurs productifs fixes ou sur les coûts fixes. Les chapitres 4 et 5 introduisent deux stratégies d'identification. Dans le chapitre 4, nous proposons une spécification de la fonction de coût qui comprend une partie fixe et une partie variable (*two-points flexible fonctionnelle form*). Les résultats empiriques basées sur cette fonction de coût suggèrent qu'une partie considérable des coûts de production est fixe, et il y a un arbitrage entre coûts fixes et coûts variables. Nous constatons également que les industries à coûts fixes élevés, produisent plus en moyenne, bénéficient d'un pouvoir de marché ainsi que de rendements d'échelle plus élevés, et sont souvent plus concentrés. Dans le chapitre 5, nous analysons le coût d'entrée sur le marché d'exportation. Notre stratégie empirique consiste à comparer les revenus des exportateurs nouvellement entrés sur le marché international avec

ceux des exportateurs déjà établis, en utilisant un modèle d'évaluation du traitement (*treatment evaluation model*). Nous constatons que les coûts d'entrée existent dans le secteur manufacturier français et que les entreprises les plus productives sont en mesure d'entrer sur le marché avec un moindre coût d'entrée.

Chapter 8

General references

Acemoglu, D., "Labor-and capital-augmenting technical change," *Journal of the European Economic Association* 1 (2003), 1–37.

Acemoglu, D. and R. Shimer, "Wage and Technology Dispersion," *Review of Economic Studies* 67 (2000), 585–607.

Ackerberg, D., C. L. Benkard, S. Berry and A. Pakes, "Econometric tools for analyzing market outcomes," in *Handbook of Econometrics*volume 6 (Elsevier, 2007), 4171–4276.

Ackerberg, D., K. Caves and G. Frazer, "Structural identification of production functions," *MPRA Paper 38349* (2006).

Adcock, R. J., "A problem in least squares," *The Analyst* 5 (1878), 53–54.

Aghion, P. and P. Howitt, "A model of growth through creative destruction," *Econometrica* 60 (1992), 323–51.

Anderson, T. W. and H. Rubin, "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics* 20 (1949), 46–63.

Antràs, P., "Is the U.S. aggregate production function Cobb-Douglas? new estimates of the elasticity of substitution," *The BE Journal of Macroeconomics* 4 (2004), 1–36.

Arellano, M. and S. Bond, "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies* 58 (1991), 277–297.

Arrow, K., "Economic welfare and the allocation of resources for invention," in *The rate and direction of inventive activity: Economic and social factors* (NBER, 1962), 609–626.

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

Baldwin, R., "Hysteresis in import prices: the beachhead effect," *American Economic Review* 78 (1988), 773–785.

———, "Sunk-Cost Hysteresis," *NBER Working Paper 2911* (1989).

Baldwin, R. and P. Krugman, "Persistent trade effects of large exchange rate shocks," *Quarterly Journal of Economics* 104 (1989), 635–654.

Bartelsman, E. J., "Of empty boxes: Returns to scale revisited," *Economics Letters* 49 (1995), 59–67.

Bartlesman, E. and W. B. Gray, "The NBER manufacturing productivity database," Technical Report, NBER, 1996.

Basu, S., "Returns to scale measurement," in S. N. Durlauf and L. E. Blume, eds., *The New Palgrave Dictionary of Economics* (Basingstoke: Palgrave Macmillan, 2008).

Basu, S. and J. G. Fernald, "Returns to scale in US production: estimates and implications," *Journal of Political Economy* 105 (1997), 249–283.

Baumol, W., J. Panzar and R. Willig, *Contestable markets and the theory of industry structure* (Harcourt Brace Jovanovich, 1988).

Baumol, W. J. and R. D. Willig, "Fixed costs, sunk costs, entry barriers, and sustainability of monopoly," *Quarterly Journal of Economics* 96 (1981), 405–431.

Bernard, A., J. Eaton, J. Jensen and S. Kortum, "Plants and productivity in international trade," *American Economic Review* 93 (2003), 1268–1290.

Bernard, A. and J. Wagner, "Export entry and exit by German firms," *Review of World Economics* 137 (2001), 105–123.

Bernard, A. B. and J. B. Jensen, "Exceptional exporter performance: cause, effect, or both?," *Journal of International Economics* 47 (1999), 1–25.

———, "Why some firms export," *Review of Economics and Statistics* 86 (2004), 561–569.

Bernard, A. B., S. J. Redding and S. P. K., "Comparative advantage and heterogeneous firms," *Review of Economic Studies* 74 (2007), 31–66.

Berndt, E. R., "Reconciling alternative estimates of the elasticity of substitution," *Review of Economics and Statistics* 58 (1976), 59–68.

Berry, S. and P. Reiss, "Empirical Models of Entry and Market Structure," in M. Armstrong and R. Porter, eds., *Handbook of Industrial Organization*volume 3, chapter 29 (Elsevier, 2007), 1845–1886.

Blackorby, C. and W. Schworm, "The structure of economies with aggregate measures of capital: a complete characterization," *Review of Economic Studies* 51 (1984), 633–650.

———, "The existence of input and output aggregates in aggregate production functions," *Econometrica* 56 (1988), 613–43.

Blundell, R. and J.-M. Robin, "Latent separability: grouping goods without weak separability," *Econometrica* 68 (2000), 53–84.

Bresnahan, T. F., "The oligopoly solution concept is identified," *Economics Letters* 10 (1982), 87–92.

Browning, M. J., "Necessary and sufficient conditions for conditional cost functions," *Econometrica* 51 (1983), 851–56.

Cabral, L., "Technology uncertainty, sunk costs, and industry shakeout," *Industrial and Corporate Change* 21 (2012), 539–552.

Cameron, A. and P. Trivedi, *Microeconometrics: Methods and Applications* (Cambridge University Press, 2005).

Campa, J. M., "Exchange rates and trade: How important is hysteresis in trade?," *European Economic Review* 48 (2004), 527–548.

Caves, D. W., L. R. Christensen and J. A. Swanson, "Productivity growth, scale economies, and capacity utilization in US railroads, 1955-74," *American Economic Review* 71 (1981), 994–1002.

Chambers, R. G., *Applied production analysis: a dual approach* (Cambridge University Press, 1988).

Chaney, T., "Distorted gravity: the intensive and extensive margins of international trade," *American Economic Review* 98 (2008), 1707–1721.

Chen, X., "Increasing returns to scale in U.S. manufacturing industries: evidence from direct and reverse regression," *BETA Working Paper 2011-11* (2011).

———, "Estimation of the CES production function with biased technical change:a control function approach," *BETA Working Paper 2012-20* (2012).

Chen, X. and B. Koebel, "Fixed cost, variable cost and the markup," *BETA Working Paper 2013-13* (2013).

Chen, X. and F. Olland, "Self-selection into export market: Does productivity affect the entry barriers?," *mimeo* (2013).

Chesher, A., "Identification in nonseparable models," *Econometrica* 71 (2003), 1405–1441.

Chirinko, R. S., "σ: the long and short of it," *Journal of Macroeconomics* 30 (2008), 671–686.

Christensen, L. R., D. W. Jorgenson and L. J. Lau, "Conjugate duality and the transcendental logarithmic production function," *Econometrica* 39 (1971), 255–256.

———, "Transcendental logarithmic production frontiers," *Review of Economics and Statistics* 55 (1973), 28–45.

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

Colander, D., "On the treatment of fixed and sunk costs in the principles textbooks," *Journal of Economic Education* 35 (2004), 360–364.

Costantini, J. and M. Melitz, "The dynamics of firm-level adjustment to trade liberalization," in *The organization of firms in a global economy* (Harvard University Press, 2007), 107–141.

Crozet, M. and F. Trionfetti, "Comparative Advantage and Within-Industry Firms Performance," *CEPII Working Paper 2011-01* (2011).

Das, S., M. J. Roberts and J. R. Tybout, "Market entry costs, producer heterogeneity, and export dynamics," *Econometrica* 75 (2007), 837–873.

Davidson, R. and J. G. MacKinnon, *Estimation and inference in econometrics* (Oxford University Press, 1993).

De Loecker, J., "Do exports generate higher productivity? Evidence from Slovenia," *Journal of International Economics* 73 (2007), 69–98.

———, "Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity," *Econometrica* 79 (2011), 1407–1451.

———, "Detecting Learning by Exporting," *American Economic Journal: Microeconomics* 5 (2013), 1–21.

Dehez, P., J. H. Dreze and T. Suzuki, "Imperfect competition a la Negishi, also with fixed costs," *Journal of Mathematical Economics* 39 (2003), 219–237.

Diewert, W., "Duality approaches to microeconomic theory," in K. J. Arrow and M. Intriligator, eds., *Handbook of Mathematical Economics*volume 2 of *Handbook of Mathematical Economics*, chapter 12 (Elsevier, 1993), 535–599.

Diewert, W. E., "An application of the Shephard duality theorem: a generalized Leontief production function," *Journal of Political Economy* 79 (1971), 481–507.

———, "Exact and superlative index numbers," *Journal of Econometrics* 4 (1976), 115–145.

———, "Cost functions," in S. N. Durlauf and L. E. Blume, eds., *The New Palgrave Dictionary of Economics* (Basingstoke: Palgrave Macmillan, 2008).

Diewert, W. E. and K. J. Fox, "On the estimation of returns to scale, technical progress and monopolistic markups," *Journal of Econometrics* 145 (2008), 174–193.

Diewert, W. E. and T. J. Wales, "Flexible Functional Forms and Global Curvature Conditions," *Econometrica* 55 (1987), 43–68.

Dixit, A., "Entry and exit decisions under uncertainty," *Journal of Political Economy* 97 (1989a), 620–638.

———, "Hysteresis, import penetration, and exchange rate pass-through," *Quarterly Journal of Economics* 104 (1989b), 205–228.

Donald, S. G. and W. K. Newey, "Choosing the number of instruments," *Econometrica* 69 (2001), 1161–1191.

Eaton, J., M. Eslava, M. Kugler and J. Tybout, "Export dynamics in colombia: firm-level evidence," *NBER Working Paper 13531* (2007).

Eaton, J., S. Kortum and F. Kramarz, "An Anatomy of International Trade: Evidence From French Firms," *Econometrica* 79 (2011), 1453–1498.

Farmer, R. E. and J.-T. Guo, "Real business cycles and the animal spirits hypothesis," *Journal of Economic Theory* 63 (1994), 42–72.

Frisch, R., "Editor's Note," *Econometrica* 1 (1933), 1–4.

———, "The principle of substitution: an example of its application in the chocolate industri," *Nordisk Tidsskrift for Teknisk Okonomi* 1 (1935), 12–27.

Fuller, W. A., "Some properties of a modification of the limited information estimator," *Econometrica* 45 (1977), 939–53.

Fuss, M. A., "The structure of technology over time: a model for testing the "Putty-Clay" Hypothesis," *Econometrica* 45 (1977), 1797–1821.

Gillard, J., "An historical overview of linear regression with errors in both variables," *Cardiff University School of Mathematics Technical Report* (2006).

Gladden, T. and C. Taber, "The relationship between wage growth and wage levels," *Journal of Applied Econometrics* 24 (2009), 914–932.

Golub, G. H. and C. F. Van Loan, "An analysis of the total least squares problem," *SIAM Journal on Numerical Analysis* 17 (1980), 883–893.

Gorman, W., *Separability and Aggregation, Collected Works of WM Gorman, Volume I, edited by C. Blackorby and AF Shorrocks* (Clarendon Press: Oxford, 1995).

Griliches, Z. and J. Mairesse, "Production functions: the search for identification," *NBER Workig Paper 5067* (1995).

Hahn, J. and J. Hausman, "A new specification test for the validity of instrumental variables," *Econometrica* 70 (2002), 163–189.

Hahn, J., Y. Hu and G. Ridder, "Instrumental variable estimation of nonlinear models with nonclassical measurement error using control variates," *mimeo* (2008).

Hall, R. E., "The Relation between price and marginal cost in US industry," *Journal of Political Economy* 96 (1988), 921–47.

———, "Invariance properties of Solow's productivity residual," *NBER Working Paper 3034* (1991).

Heckman, J. J., "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *NBER Working Paper 2861* (1989).

Heckman, J. J. and V. J. Hotz, "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training," *Journal of the American Statistical Association* 84 (1989), 862–874.

Heckman, J. J. and R. Robb, "Alternative methods for evaluating the impact of interventions: An overview," *Journal of Econometrics* 30 (1985), 239–267.

Heckman, J. J. and E. J. Vytlacil, "Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation," in J. Heckman and E. Leamer, eds., *Handbook of Econometrics*volume 6, chapter 70 (Elsevier, 2007).

Helpman, E., S. Yeaple and M. Melitz, "Export versus FDI with heterogeneous firms," *American Economic Review* 94 (2004), 300–316.

Hicks, J. R., *The Theory of Wages* (Macmillan London, 1932).

———, *Value and capital*, volume 2 (Clarendon Press Oxford, 1946).

Hong, J., "Nonparametric Identification and Estimation of Production Functions Using Control Function Approaches to Endogeneity," *mimeo* (2008).

Hotelling, H., "Edgeworth's taxation paradox and the nature of demand and supply functions," *Journal of Political Economy* 40 (1932), 577–616.

Imbens, G. W., "Nonadditive models with endogenous regressors," *Econometric Society Monographs* 43 (2007), 17.

Imbens, G. W. and W. K. Newey, "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica* 77 (2009), 1481–1512.

Imbens, G. W. and J. M. Wooldridge, "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature* 47 (2009), 5–86.

Javorcik, B. S., "Does foreign direct investment increase the productivity of domestic firms? In search of spillovers through backward linkages," *American Economic Review* 94 (2004), 605–627.

Jones, C. I., "Growth and ideas," volume 1 (Elsevier, 2005), 1063–1111.

Jorgenson, D. W., "Econometric methods for modeling producer behavior," in Z. Griliches and M. D. Intriligator, eds., *Handbook of Econometrics*chapter 31 (Elsevier, 1986), 1841–1915.

Judge, G. G., W. E. Griffiths, R. C. Hill and T.-C. Lee, *The Theory and Practice of Econometrics* (New York Wiley, 1980).

Kendall, M. and A. Stuart, *The Advanced Theory of Statistics* (New York Hafner Publishing, 1973).

Klump, R., P. McAdam and A. Willman, "Factor substitution and factor-augmenting technical progress in the United States: a normalized supply-side system approach," *Review of Economics and Statistics* 89 (2007), 183–192.

———, "The normalized CES production function: theory and empirics," *Journal of Economic Surveys* 26 (2012), 769–799.

Kmenta, J., "On estimation of the CES production function," *International Economic Review* 8 (1967), 180–189.

Koebel, B., M. Falk and F. Laisney, "Imposing and Testing Curvature Conditions on a Box-Cox Cost Function," *Journal of Business and Economic Statistics* 21 (2003), 319–35.

Koebel, B. and F. Laisney, "The Aggregate Le Chatelier Samuelson principle with Cournot competition," *ZEW-Centre for European Economic Research Discussion Paper 10-009* (2010).

Konings, J. and H. Vandenbussche, "Heterogeneous responses of firms to trade protection," *Journal of International Economics* 76 (2008), 371–383.

Krugman, P., *Exchange-Rate Instability*, The Lionel Robbins Lectures (Mit Press, 1989).

Krugman, P. R., "Increasing returns, monopolistic competition, and international trade," *Journal of International Economics* 9 (1979), 469–479.

Kumbhakar, S. C. and E. G. Tsionas, "Stochastic error specification in primal and dual production systems," *Journal of Applied Econometrics* 26 (2011), 270–297.

Lau, L. J., "A characterization of the normalized restricted profit function," *Journal of Economic Theory* 12 (1976), 131–163.

———, "Testing and imposing monoticity, convexity and quasi-convexity constraints," in M. Fuss and D. McFadden, eds., *Production economics: a dual approach to theory and applications*volume 1 (North Holland, 1978), 409–453.

———, "On identifying the degree of competitiveness from industry price and output data," *Economics Letters* 10 (1982), 93–99.

León-Ledesma, M. A., P. McAdam and A. Willman, "Identifying the elasticity of substitution with biased technical change," *American Economic Review* 100 (2010), 1330–1357.

Leontief, W., "Introduction to a theory of the internal structure of functional relationships," *Econometrica* 15 (1947), 361–373.

Levinsohn, J. and A. Petrin, "Estimating production functions using inputs to control for unobservables," *Review of Economic Studies* 70 (2003), 317–341.

Lewbel, A., "Direct utility functions for the PIGL, PIGLOG, quadratic logarithmic, and Gorman polar form," *mimeo, Department of Economics, Brandeis University.* (1998).

Madansky, A., "The fitting of straight lines when both variables are subject to error," *Journal of the American Statistical Association* 54 (1959), 173–205.

Malley, J. and V. A. Muscatelli, "Business cycles and productivity growth: are temporary downturns productive or wasteful?," *Research in Economics* 53 (1999), 337–364.

Mariano, R. S. and T. Sawa, "The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included endogenous variables," *Journal of the American Statistical Association* 67 (1972), 159–163.

Mas-Colell, A., M. Whinston and J. Green, *Microeconomic Theory*, Oxford student edition (Oxford University Press, 1995).

Melitz, M. J., "The impact of trade on intra-industry reallocations and aggregate industry productivity," *Econometrica* 71 (2003), 1695–1725.

Morrison, C. J., "Quasi-fixed inputs in U.S. and Japanese manufacturing: a generalized leontief restricted cost function approach," *Review of Economics and Statistics* 70 (1988), 275–87.

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

———, "On the pooling of time series and cross section data," *Econometrica* 46 (1978), 69–85.

Murphy, K. M., A. Shleifer and R. W. Vishny, "Industrialization and the big push," *Journal of Political Economy* 97 (1989), 1003–1026.

Nagar, A., "The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations," *Econometrica* 27 (1959), 575–595.

Nerlove, M., "Recent empirical studies of the CES and related production functions," in *The theory and empirical analysis of production* (Columbia University Press, 1967), 55–136.

Newey, W. K., "Flexible simulated moment estimation of nonlinear errors-in-variables models," *Review of Economics and Statistics* 83 (2001), 616–627.

Newey, W. K. and K. D. West, "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica* 55 (1987), 703–708.

Olley, G. S. and A. Pakes, "The dynamics of productivity in the telecommunications equipment industry," *Econometrica* 64 (1996), 1263–1297.

Pindyck, R. S. and J. J. Rotemberg, "Dynamic factor demands and the effects of energy price shocks," *American Economic Review* 73 (1983), 1066–1079.

Roberts, M. J. and J. R. Tybout, "The decision to export in colombia: an empirical model of entry with sunk costs," *American Economic Review* 87 (1997), 545–564.

Robinson, P. M., "Root-N-Consistent semiparametric regression," *Econometrica* 56 (1988), 931–954.

Rosenbaum, P. R. and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70 (1983), 41–55.

Samuelson, P. A., "Structure of a minimum equilibrium system," in *Essays in Economics and Econometrics: A Volume in Honor of Harold Hotelling* (University of North Carolina Press, 1960), 1–33.

Sato, K., "A two-level constant-elasticity-of-substitution production function," *Review of Economic Studies* 34 (1967), 201–218.

Schmidt, P., *Econometrics*, Statistics, textbooks and monographs (M. Dekker, 1976).

Schumpeter, J., "The common sense of econometrics," *Econometrica* 1 (1933), 5–12.

Shea, J., "Instrument relevance in multivariate linear models: a simple measure," *Review of Economics and Statistics* 79 (1997), 348–352.

Shephard, R. W., *Theory of cost and production functions*, volume 4 (Princeton University Press, 1970).

Solow, R., "Capital, Labor, and Income in Manufacturing," in *The Behavior Of Income Shares: Selected Theoretical and Empirical Issues*, NBER Chapters (National Bureau of Economic Research, Inc, 1964), 101–142.

Staiger, D. and J. H. Stock, "Instrumental variables regression with weak instruments," *Econometrica* 65 (1997), 557–586.

Stefanski, L., "Measurement error models," *Journal of the American Statistical Association* 95 (2000), 1353–1358.

Stock, J. and M. Yogo, "Testing for weak instruments in linear IV regression," *NBER Working Paper 0284* (2002).

Stock, J. H., J. H. Wright and M. Yogo, "A Survey of weak instruments and weak identification in generalized method of moments," *Journal of Business and Economic Statistics* 20 (2002), 518–29.

Sutton, J., "Market Structure: Theory and Evidence," in M. Armstrong and R. Porter, eds., *Handbook of Industrial Organization*volume 3, chapter 35 (Elsevier, 2007), 2301–2368.

Swamy, P. A. V. B., "Efficient inference in a random coefficient regression model," *Econometrica* 38 (1970), 311–23.

Syverson, C., "What determines productivity?," *Journal of Economic Literature* 49 (2011), 326–65.

Theil, H., "Specification errors and the estimation of economic relationships," *Revue de l'Institut International de Statistique* 25 (1957), 41–51.

Thursby, J. G. and C. A. K. Lovell, "An investigation of the Kmenta approximation to the CES function," *International Economic Review* 19 (1978), 363–77.

Uzawa, H., "Production functions with constant elasticities of substitution," *Review of Economic Studies* 29 (1962), 291–299.

Van Beveren, I., "Total factor productivity estimation: a practical review," *Journal of Economic Surveys* 26 (2012), 98–128.

Van Biesebroeck, J., "Exporting raises productivity in sub-Saharan African manufacturing firms," *Journal of International Economics* 67 (2005), 373–391.

Viner, J., "Costs Curves and Supply Curves," *Zeitschrift fur Nationalokonomie* 3 (1931), 23–46.

Wald, A., "The fitting of straight lines if both variables are subject to error," *Annals of Mathematical Statistics* 11 (1940), 284–300.

Wang, X. H. and B. Z. Yang, "Fixed and sunk costs revisited," *Journal of Economic Education* 32 (2001), 178–185.

———, "On the treatment of fixed and sunk costs in principles textbooks: A comment and a reply," *Journal of Economic Education* 35 (2004), 365–369.

Wansbeek, T. J. and E. Meijer, *Measurement error and latent variables in econometrics*, volume 37 (Elsevier Amsterdam, 2000).

Wooldridge, J. M., *Econometric Analysis of Cross Section and Panel Data* (MIT press, 2002).

———, "On estimating firm-level production functions using proxy variables to control for unobservables," *Economics Letters* 104 (2009), 112–114.

UNIVERSITÉ DE STRASBOURG

Xi CHEN

Technologie, productivité et coûts fixes : quatre essais sur l'analyse de la production appliquée

Résumé

Cette thèse se compose de quatre essais sur l'analyse de la production appliquée, avec un accent mis sur la technologie, la productivité et les coûts fixes. L'objectif de cette thèse est d'identifier les limites de la modélisation du comportement de producteurs, et de proposer certaines améliorations. Dans cette thèse, nous avons comparé les principales spécifications empiriques et les méthodes statistiques qui ont été utilisées dans l'analyse de la production, et souligné leurs implications pour l'estimation des paramètres technologiques. Nous avons étudié les causes et les remèdes au problème d'endogénéité dans le cadre de l'estimation de la fonction de production. Cette thèse a également abordé un aspect important de la production et pourtant largement négligé dans la littérature : les coûts fixes. Ce travail a contribué à la définition et à la caractérisation des coûts fixes. Nous avons développé des stratégies d'estimation du coût fixe, et montré que le coût fixe a un impact significatif sur la politique de prix, les rendements d'échelle et les exportations.

Résumé en anglais

This thesis consists of four essays on applied production analysis, with a focus on technology, productivity and fixed costs. The aim of this thesis is to identify some limitations of recent contributions to production behavior modeling, and to propose improvements. In this dissertation, I compared different empirical specifications and statistical methods which have often been used in production analysis, and pointed out their implications for estimating technology parameters. I studied the causes and cures of endogeneity problems in the context of production analysis. This thesis also addressed the important but neglected issue of fixed costs. This work defined and characterized the fixed cost, and developed empirical strategies to estimate the fixed cost using the standard production database. Empirical evidence suggests that the fixed cost is significant and has profound impacts on producer's behavior in terms of price setting, returns to scale and exports.