ECONOMETRIC ANALYSIS OF FIRM-LEVEL PRODUCTION DATA

By

JOHN M. KEALEY, B.A., M.A., M.A.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2016) McMaster University
(Economics) Hamilton, Ontario

Title: Econometric analysis of firm-level production data

Author: John M. Kealey
B.A. (University of Victoria)
M.A. (Dalhousie University)
M.A. (Carleton University)

Supervisor: Professor Jeffrey Racine

Number of Pages: xiv, 163

Abstract

In this dissertation, I explore a variety of methods for the econometric analysis of firm-level production data. Three distinct approaches are considered, namely i) proxy variable methods of controlling for unobservable productivity, ii) data envelopment techniques for estimating the boundary of a production set, and iii) stochastic frontier methods for estimating the productive inefficiency of firms. Much of the focus is on semiparametric and nonparametric estimators that allow for a highly flexible specification of the function that relates input combinations to output quantities.

After a brief introductory chapter that outlines the theme and objectives of this dissertation, in the second chapter, I propose a semiparametric method of estimating a Cobb-Douglas model of firm-level production in which the elasticity coefficients can be functions of both continuous and discrete predictors. I show that the varying-coefficient method is better able to reflect the hypothesized relationship between factor elasticities and their corresponding input expenditure shares than the constant-coefficient partially linear alternative. Using plant-level data from the Colombian manufacturing sector, I provide an empirical example in which the elasticity of output with respect to capital and labour can vary by industry and across different time periods. In this setting, the contribution of unskilled labour to final output diminishes over time, which

brings about an overall decline in estimated returns to scale.

In the third chapter, I propose a robust nonparametric estimation procedure for deterministic frontier models with a count output variable. It exhibits minimal sensitivity to outliers without resorting to aggressive data trimming that tends to have an overreaching effect on non-extreme-valued observations. The estimator's favourable performance is primarily attributable to its use of a novel trimming parameter selection routine that combines k-means and hierarchical clustering techniques. Evidence from a Monte Carlo experiment suggests that both the nonsmooth and the smooth versions of the proposed estimator give rise to a better fit of the frontier function than existing robust data envelopment methods. I conclude with an empirical example that uses historical patent count data from the U.S. manufacturing sector to estimate the efficiency of firm-level R & D spending.

In the fourth chapter, I outline a nonparametric estimation procedure for stochastic production frontier models for panel data, making it possible to decompose firm-level inefficiency into a persistent and a time-varying component. Existing approaches tend to rely on a simple linear specification of the frontier function, but this can give rise to imprecise estimates of i) the frontier itself, ii) factor elasticities, and iii) firm-level inefficiency. In contrast, the method that I propose dispenses with parametric assumptions vis-a-vis the functional form of the frontier and the distribution of inefficiency, thereby avoiding some of the potentially adverse consequences of model misspecification. It is shown that the kernel-based method is better able to account for heterogeneity in production technology that exists across firms and over time. A test for first-order stochastic dominance suggests that the nonparametric framework often

yields estimates of firm-level inefficiency that are uniformly lower than those obtained by following the parametric approach.

In the fifth chapter, my co-authors and I consider whether a fairly well-established empirical relationship between liberalized trade and firm productivity growth is sensitive to the choice of an identification strategy for production function estimation. We estimate the productivity of Colombian manufacturing plants using the methods of Levinsohn and Petrin (2003), Ackerberg, Caves, and Frazer (2006), and Gandhi, Navarro, and Rivers (2016), and at times come to surprisingly different conclusions about the country's experience with trade policy reform during the 1980s. Results from a quantile regression model and a productivity growth decomposition exercise tend to vary as we experiment with different specifications of the production function. Research that is concerned with the short and medium-term impact of trade liberalization on domestic manufacturing industries should therefore pay close attention to issues of robustness to alternative strategies for estimating the productivity of firms. Finally, in the sixth chapter, I recapitulate the central methodological insights that are offered in this dissertation, and provide some concluding remarks that sum up the overall implications of my research.

Acknowledgements

I would like to thank my dissertation supervisor, Professor Jeffrey Racine, whose commitment to academic excellence has been a source of great inspiration throughout my doctoral studies. As an extremely talented researcher with a wonderful sense of humour, he has been an ideal intellectual role model, and I consider myself very fortunate to have benefitted from his personal and professional guidance. I am also grateful for the input and encouragement that I have received from other faculty in the Department of Economics at McMaster University, including Pau S. Pujolàs, César Sosa-Padilla, Mike Veall, Seungjin Han, and Catherine Kuff. I thank my fellow PhD students Luc Clair, Sean Sexton, Waheed Olagunju, and Stephanie Houle, whose friendship and scholarly collaboration have left me with many positive memories of graduate school. In addition, I gratefully acknowledge financial support that was provided by the Ontario Graduate Scholarship program in 2013-2014 and the Social Sciences and Humanities Research Council of Canada in 2014-2016.

Lastly and most importantly, I would like to thank my parents who have made many personal and financial sacrifices to give me the enduring gift of a higher education. None of my academic achievements would have been possible without their love and emotional support, and so I dedicate this thesis to them.

Contents

List of Tables

List of Figures

Declaration of academic achievement

I am the sole author of Chapters 1-4 and 6. Chapter 5 is co-authored with Pau S. Pujolàs and César Sosa-Padilla, who devised its conceptual framework and overall structure. I was responsible for the data cleaning, statistical analysis/estimation, and the written preparation of the chapter. The raw data that underlies the analysis in Chapters 2, 4, and 5 was originally put together by Mark J. Roberts and James R. Tybout, and was made available to me by Kim J. Ruhl and Pau S. Pujolàs.

Chapter 1

Introduction

Econometric models of firm-level production are one of the foundational tools of modern empirical research in economics. While neoclassical production theory provides a rich analytical framework to better understand the input and output decisions of firms, putting its propositions into practical use with real-world data presents a number of interesting opportunities and challenges for the applied statistician. Hence this dissertation explores a variety of methods for the econometric analysis of firm-level production data, with a particular focus on semiparametric and nonparametric approaches. It offers new methodological insights pertaining to the estimation of the most salient aspects of firms' production technology, while generally avoiding parametric assumptions about the nature of the statistical relationship between input combinations and output quantities. Each of the empirical techniques that is proposed is typically illustrated by means of a Monte Carlo simulation in which its finite-sample performance is contrasted with that of alternative approaches that have been adopted previously in the economics literature. In addition, each of the econometric routines that is presented is reinforced by an applied example that

involves firm-level data from the manufacturing sector.

There exist three broad categories of models of firm-level production, each of which has its own dedicated chapter in this thesis. *Control function* methods introduced by Olley and Pakes (1996) and Levinsohn and Petrin (2003) use proxy variables to control for firm productivity which is modeled as an unobservable residual from an equation that relates input combinations to output quantities. This type of approach is ideally-suited for models of production that are subject to simultaneity bias whereby the unobservable productivity of firms tends to influence their input decisions, which can lead to biased estimates of the marginal importance of certain factors of production (Marschak and Andrews, 1944; Griliches and Mairesse, 1995). *Deterministic frontier analysis* originates in the work of Farrell (1957), Charnes, Cooper, and Rhodes (1978), and Deprins, Simar, and Tulkens (1984), and is concerned with estimation of the boundary of a production set, which is defined as the collection of technologically-feasible input-output combinations. In this setting, the objective is to identify the maximum attainable output quantity for any particular choice of inputs, and then to compute the productive inefficiency of each firm as its output shortfall vis-à-vis the frontier. In practice, data envelopment methods can be used to estimate the boundary of the production set for a sample of firms. On the other hand, a *stochastic frontier* doesn't necessarily need to envelop every observation in the data. In this type of model, which was first explored by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), the distance that separates a firm's realized output and the production frontier is not uniquely attributable to inefficiency; it is assumed that part of the deviation stems from unexplained variation in

output that might be due to either random noise or measurement error. Thus, a central challenge is to extract the inefficiency component of the convoluted error term after the frontier function has been estimated.

If there is a single unifying theme to be found in the four chapters that make up the main body of this dissertation, it is the importance of *statistically robust* estimation of models of firm-level production. From this perspective, robust econometric methods for the analysis of firm data tend to satisfy two criteria, among others: first, they are applicable to a broad range of specifications of the production model, not just one. Parametric approaches do not, in general, meet this requirement because they rely on a somewhat rigid characterization of the production set, factor elasticities, returns to scale, and the productive efficiency of firms. Hence it is often preferable to adopt a more robust semiparametric or nonparametric framework that allows for greater flexibility in these and other respects. The second robustness criterion pertains to an estimator's sensitivity to outliers. Extreme-valued observations can be quite problematic in the context of production frontier modelling, since the objective is to estimate a function that envelops the set of feasible output quantities for any given input combination. Given that traditional methods of frontier estimation are unduly influenced by outliers, it can be quite beneficial to follow a more robust approach that is resistant to anomalies in the data. While these two robustness criteria for models of firm-level production have received quite a bit of attention in the econometrics literature, they are not always given adequate consideration in applied research settings. In this dissertation, I draw attention to some of the aberrant outcomes that one might encounter when dealing with models that are built on erroneous functional

form assumptions or with data that is compromised by measurement error, and show how these complications can be avoided when conducting empirical research on the productivity of firms.

In Chapter 2, I propose a semiparametric varying coefficient estimation procedure for production functions that use a proxy variable to control for unobservable productivity. I draw attention to a difficulty that can arise when applying the standard framework of Cobb and Douglas (1928) to situations where there is intra-industry heterogeneity in factor payments as a share of firm revenue. In particular, the log-linearized Cobb-Douglas production function tends to treat the elasticity of output with respect to each of the factors of production as a fixed parameter, even though the corresponding input expenditure's share of revenue is almost never the same for all firms. Thus, the traditional parametric approach cannot be considered robust to different specifications of the factor elasticities in the model of firm-level production. In contrast, the semiparametric estimator that is offered as an alternative treats factor elasticities and returns to scale as flexible functions of mixed data type predictors that vary across firms, industries, or time periods, which yields greater consistency with standard theories of firm behaviour as it relates to input decisions. This estimation procedure is especially useful in a panel data context, where the researcher might suspect that the production technology of a certain industry has undergone significant changes over time. Under this scenario, one can apply local weighting to the elasticity coefficient estimator based on the value of a discrete or a continuous time variable, which will allow for intertemporal heterogeneity in the function that relates factors of production with firm-level output.

In Chapter 3, I outline a framework for robust nonparametric estimation of deterministic production frontiers for count data. As mentioned earlier, frontier models of the deterministic variety assume that all observed input-output combinations in a sample of firms are enveloped by the boundary of the production set, and consequently, they are often not robust to the presence of outliers in the data. Building on previous work by Cazals, Florens, and Simar (2002), Aragon, Daouia, and Thomas-Agnan (2005), and Martins-Filho and Yao (2008), who make use of data trimming procedures to either reduce or eliminate the standard frontier estimator's sensitivity to extreme-valued observations, I propose a novel approach to data trimming parameter selection that relies on hierarchical clustering as a means of distinguishing robust frontier estimates from those that are driven by a handful of outliers. The robust production frontier is based on the distribution of order-one conditional output quantile estimates for a large number of resampled subsets of observations in the data. One of the interesting features that distinguishes this particular methodological discussion from the ones that appear in the other chapters is its focus on the special case of an output variable that is expressed as a discrete count, rather than a continuous quantity. The analysis of count data in a frontier modelling context offers many interesting possibilities, since there are a number of industries where it is more realistic for the empirical researcher to assume that the output of firms has been drawn from a discrete distribution.

In Chapter 4, I delineate a fully nonparametric method of estimating stochastic frontier models for panel data, and contrast its characterization of the output frontier and the productive efficiency of firms with that of Heshmati and Kumbhakar's (1995) linear parametric framework. The functional

form and distributional assumptions that underlie the parametric method are closely scrutinized, and it appears that model misspecification is a very real concern. For instance, as mentioned earlier, stochastic frontier estimation involves a deconvolution problem whereby the productive inefficiency of firms must be distinguished from the random noise component of the model. In a parametric or semiparametric framework, this requires assumptions about their respective densities so that they can be estimated using a maximum likelihood approach once an expression for the convoluted error term has been formulated. Amsler, Schmidt, and Wang (2011) propose two goodness-of-fit tests that can help determine whether these density functions have been correctly specified, and in the context of the example that is considered in the present study, two of the most common distributional assumptions that are found in the literature are frequently rejected by the data. Therefore, the parametric stochastic frontier model for panel data isn't really robust to alternative specifications of the convoluted error's density, which suggests that a different methodology for the estimation of firms' productive inefficiency might be preferred.

Chapter 5, which is co-authored with Pau S. Pujolàs and César Sosa-Padilla, sheds light on an empirical inconsistency that can complicate analyses of productivity growth during periods of policy reform. We consider three separate identification strategies for the estimation of production functions that have been proposed in the literature - the "control function" methods of Levinsohn and Petrin (2003) and Ackerberg, Caves, and Frazer (2006), and the more recent nonparametric approach of Gandhi, Navarro, and Rivers (2016) - and assess whether our estimates of firm productivity exhibited similar dynamic

behaviour following a trade policy liberalization episode in Colombia during the 1980s. The empirical analysis is motivated by the theoretical propositions of Melitz (2003), who shows that opening up to trade strengthens productivity overall via reallocation of resources from inefficient to efficient firms and market exit (entry) on the part of unproductive (productive) firms. Surprisingly, we find that our conclusions about how these different channels contributed to productivity growth after trade policy reform are quite dependent on our chosen identification strategy for estimation of each firm's production function. This suggests that in applied research, if the objective is to study how firms or industries might have responded to a change in the policy environment, it is important to verify whether one's results are robust to alternative specifications of the production-related variables that appear in one's model.

Securing access to reliable microdata is one of the most significant challenges faced by academic researchers who are interested in studying productivity at the firm-level. Detailed information on firms' investment in capital goods, employment of different forms of labour, and consumption of intermediate inputs such as raw materials and energy is generally only collected by means of a government-sponsored industrial census, and this data is almost never made available to the public at large. Fortunately, given that this dissertation's primary focus is on the pursuit of methodological rigour rather than uncovering fundamental truths about a particular industry in a single geographic region,[1] using the most up-to-date proprietary/confidential firm-level datasets to illustrate new econometric techniques offers little in terms of added

[1]While the focus of Chapter 5 is methodological in the sense that it highlights how alternative specifications of the production function can lead to different conclusions about whether a policy reform has encouraged productivity growth, it may nevertheless be considered an exception to this claim.

value. The empirical example at the end of Chapter 3 is based on patent count data from the U.S. manufacturing sector in 1975-1979 that first appeared in Hall, Griliches, and Hausman (1986), and that is now publicly-available as part of a computational exercise in the textbook of Cameron and Trivedi (2013). Meanwhile, the econometric methods that are explored in Chapters 2, 4, and 5 are implemented using data from a 1981-1991 census of manufacturing plants in Colombia that was first put together by Mark J. Roberts and James R. Tybout, and that has been made available to me by Kim J. Ruhl and Pau S. Pujolàs. The reader should be aware that there is substantial overlap in the preliminary data summaries that are provided in these three chapters of the thesis.

The semiparametric and nonparametric estimation strategies that are described in Chapters 2 through 4 make heavy use of kernel methods. Hayfield and Racine's (2008) `np` package and Racine and Nie's (2014) `crs` package in the R statistical computing environment are an indispensable resource for this type of analysis, insofar as they provide user-friendly routines for carrying out locally-weighted regression, kernel estimation of conditional distribution functions, and consistent hypothesis testing. The `np` package's generalized kernel summation function is also an invaluable tool, to the extent that it greatly facilitates programming of new semiparametric and nonparametric estimation procedures that do not have an already-defined command in R. In addition, wherever the processing capabilities of a consumer laptop have been insufficient to perform this project's most computationally-intensive empirical tasks, the resources of the Shared Hierarchical Academic Research Computing Network (SHARCNET) have made these workloads a lot more manageable.

The R code that is used to implement the various estimation routines that are outlined in this dissertation is available from the author upon request.

References

Ackerberg, D., K. Caves, and G. Frazer (2006): "Structural identification of production functions," MPRA paper, University Library of Munich, Germany.

Aigner, D., C. A. K. Lovell, and P. Schmidt (1977): "Formulation and estimation of stochastic frontier production function models," *Journal of Econometrics*, 6, 21–37.

Amsler, C., P. Schmidt, and W. S. Wang (2011): "Goodness of fit tests in stochastic frontier models," *Journal of Productivity Analysis*, 35, 95–118.

Aragon, Y., A. Daouia, and C. Thomas-Agnan (2005): "Nonparametric frontier estimation: A conditional quantile-based approach," *Econometric Theory*, 21, 358–389.

Cameron, A. C. and P. K. Trivedi (2013): *Regression Analysis of Count Data*, Cambridge,UK: Cambridge University Press.

Cazals, C., J. P. Florens, and L. Simar (2002): "Nonparametric frontier estimation: A robust approach," *Journal of Econometrics*, 106, 1–25.

Charnes, A., W. W. Cooper, and E. Rhodes (1978): "Measuring the inefficiency of decision making units," *European Journal of Operational Research*, 2, 429–444.

Cobb, C. W. and P. H. Douglas (1928): "A theory of production," *American Economic Review*, 18, 139–172.

Deprins, D., L. Simar, and H. Tulkens (1984): "Measuring labor inefficiency in post offices," in *The Performance of Public Enterprises: Concepts and Measurements*, ed. by M. Marchand, P. Pestieau, and H. Tulkens, Amsterdam: North-Holland, 243–267.

Farrell, M. J. (1957): "The measurement of productive efficiency," *Journal of the Royal Statistical Society Series A*, 120, 253–290.

Gandhi, A., S. Navarro, and D. Rivers (2016): "On the identification of production functions: How heterogeneous is productivity?" Manuscript.

Griliches, Z. and J. Mairesse (1995): "Production functions: The search for identification," NBER Working Papers 5067, National Bureau of Economic Research, Inc.

Hall, B. H., Z. Griliches, and J. A. Hausman (1986): "Patents and R & D: Is there a lag?" *International Economic Review*, 27, 265–283.

Hayfield, T. and J. S. Racine (2008): "Nonparametric econometrics: The np package," *Journal of Statistical Software*, 27.

Heshmati, A. and S. C. Kumbhakar (1995): "Efficiency measurement in Swedish dairy farms: An application of rotating panel data," *American Journal of Agricultural Economics*, 77, 660–674.

Levinsohn, J. and A. Petrin (2003): "Estimating production functions using inputs to control for unobservables," *Review of Economic Studies*, 70, 317–341.

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

Martins-Filho, C. and F. Yao (2008): "A smooth nonparametric conditional quantile frontier estimator," *Journal of Econometrics*, 143, 317–333.

Meeusen, W. and J. van den Broeck (1977): "Efficiency estimation for Cobb-Douglas production functions with composed error," *International Economic Review*, 18, 435–444.

Melitz, M. J. (2003): "The impact of trade on intra-industry reallocations and aggregate industry productivity," *Econometrica*, 71, 1695–1725.

Olley, S. G. and A. Pakes (1996): "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 64, 1263–97.

Racine, J. S. and Z. Nie (2014): *crs: Categorical Regression Splines*, R package version 0.15-24.

Chapter 2

Semiparametric estimation of a Cobb-Douglas production function with varying elasticity coefficients

2.1 Introduction

Methodological developments in the estimation of firm-level production functions have featured prominently in economic research for nearly a century. The continued importance that has been attached to this area of inquiry since the seminal work of Cobb and Douglas (1928) is unsurprising given its numerous applications in both micro- and macroeconomic modelling. However, a problematic feature of the conventional Cobb-Douglas specification is its somewhat dubious characterization of factor elasticities. Specifically, factor elasticities are treated as fixed parameters that do not exhibit any intertemporal or intra-industry variation, but this is incompatible with the substantial heterogeneity of firm-level input expenditure shares that is often observed in real-world data. This chapter seeks to develop a more flexible empirical frame-

work that addresses this shortcoming.

Much of the econometric literature relating to the estimation of firm-level production functions has centered on the attenuation of 'simultaneity bias,' which can arise when firms' input decisions are influenced by unobserved productivity realizations (Marschak and Andrews, 1944; Griliches and Mairesse, 1995). Special attention has been given to dynamic panel data methods (Arellano and Bond, 1991; Blundell and Bond, 1998, 2000) and proxy variable approaches (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2006; Wooldridge, 2009). However, these empirical techniques fail to consider any variation in factor shares across firms and over time by treating their corresponding Cobb-Douglas elasticities as constant. Of course, one way to rectify this apparent contradiction is to dispense with rigid parametric specifications of firm-level input-output models. For instance, Gandhi, Navarro, and Rivers (2016) have sketched out a framework for the identification and estimation of production functions in a nonparametric setting. This avenue of research holds promise insofar as nonparametric and semiparametric econometric techniques have generally not received much attention in the literature.

The present chapter shows how semiparametric varying coefficient methods can be used to estimate a Cobb-Douglas production function with highly flexible factor elasticities, which allows the researcher to abandon some of the restrictive assumptions that have hindered previous methodological approaches. Most importantly, the proposed framework makes it possible to obtain elasticity estimates that follow similar patterns of variation as their theoretically-analogous input expenditure shares, which represents an impor-

tant innovation in the literature. Varying coefficient models have proven popular among statisticians and econometricians since they constitute somewhat of a "middle ground" between fully parametric and fully nonparametric specifications. This class of estimators originates in the work of Hastie and Tibhirani (1993) and Chen and Tsay (1993), and can be implemented using either spline methods (Huang, Wu, and Zhou, 2004; Lee, Mammen, and Park, 2012) or kernel weighting (Li, Huang, Li, and Fu, 2002; Cai and Li, 2008; Li and Racine, 2010). Many of the key theoretical developments in this area of semiparametric modelling are discussed in Park, Mammen, Lee, and Lee (2015).

The remainder of this chapter is structured as follows: Section 2.2 outlines the semiparametric estimation procedure for a Cobb-Douglas production function with varying elasticity coefficients, while Section 2.3 provides an empirical example that makes use of a plant-level dataset from the Colombian manufacturing sector. Section 2.4 presents the results of a Monte Carlo experiment that assesses the proposed estimator's finite sample performance, while Section 2.5 concludes.

2.2 Varying elasticity coefficients in a Cobb-Douglas setting

Consider a logarithmic transform of a Cobb-Douglas production function:

$$y_{it} = \mathbf{x}_{it}\beta + \mathbf{w}_{it}\gamma + v_{it} + e_{it}, \tag{2.1}$$

where y_{it} is the natural log of firm i's value-added output in period t, $\mathbf{x}_{it}$ is the natural log of a vector of quasi-fixed inputs, $\mathbf{w}_{it}$ is the natural log of a vector of

variable inputs, v_{it} is the firm's (unobservable) total factor productivity, and e_{it} is a random disturbance term whose mean conditional on current and past inputs is equal to zero. The vectors β and γ, which the econometrician would like to estimate, denote the elasticity of output with respect to the quasi-fixed and variable inputs, respectively. For ease of comparison, (2.1) is expressed in the same notation that appears in Wooldridge (2009). Of course, one of the assumptions that is implied by the functional form of (2.1) is that production technology, as reflected by the elasticity parameters β and γ, is homogeneous among all firms and across all time periods. In panel datasets that comprise firms with diverse characteristics, however, this level of inflexibility in the elasticity parameters can potentially be problematic.

Elementary microeconomic theory of the firm suggests that factor elasticities and input expenditure shares ought to be closely related. If the variable input prices that correspond to the d-dimensional vector $\mathbf{w}_{it} = (w_{1it}, ..., w_{dit})$ are denoted by $p_{w_1}, ..., p_{w_d}$, and if the price of the final good is denoted by p_y, then the first-order conditions governing the firm's input decisions yield the following identity:

$$\gamma_{w_j} = \frac{p_{w_j} W_{jit}}{p_y Y_{it}} \qquad j = 1, ..., d, \tag{2.2}$$

where the right-hand side of each of the identities in (2.2) is easily interpreted as the ratio of variable input expenditures to total revenue (lower-case and upper-case letters indicate that a variable is being expressed in log and level form, respectively). It is worthy of note that equation (2.2) points to a potential inconsistency between standard theories of firm behaviour on the one hand, and much of the econometric literature dealing with the estimation of production functions on the other. In particular, unless one assumes that the

ratios $\frac{p_{w_j} W_{jit}}{p_y Y_{it}}$ remain fixed across all plants and over all time periods, in empirical applications, it does not make much sense to estimate the elasticity coefficients $\gamma_{w_1}, ..., \gamma_{w_d}$ as constant parameters.[1] In the light of the possibility that firm-level production technology is more heterogeneous than previously assumed, it might be advisable to make a small modification to the functional form of (2.1):

$$y_{it} = \mathbf{x}_{it} \cdot \beta(\mathbf{z}_{it}) + \mathbf{w}_{it} \cdot \gamma(\mathbf{z}_{it}) + v_{it} + e_{it}, \tag{2.3}$$

where β and γ are now functions of the vector $\mathbf{z}_{it}$ that can include a mixture of continuous and discrete data.[2] Specifically, let $\mathbf{z}_{it} = (\mathbf{z}^c_{it}, \mathbf{z}^u_{it}, \mathbf{z}^o_{it}) = \left(z^c_{1,it}, ..., z^c_{q,it}, z^u_{1,it}, ..., z^u_{r,it}, z^o_{1,it}, ..., z^o_{s,it}\right)$ denote a $q + r + s$-dimensional vector that consists of q continuous, r unordered discrete, and s ordered discrete variables that encapsulate industry, time, and firm-specific characteristics.

An important consideration that has been discussed at length in the empirical industrial organization literature, and one that will affect the identification of $\beta(\mathbf{z}_{it})$ and $\gamma(\mathbf{z}_{it})$ in (2.3), is the influence that the unobservable productivity term v_{it} likely has on the firm's choice of its period-t variable input quantity $\mathbf{w}_{it}$. Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg et al. (2006), and others have shown that one way of dealing with this challenge is to use a vector of proxy variables, $\mathbf{m}_{it}$, that are observable in the data and strictly increasing in v_{it}, holding $\mathbf{x}_{it}$ fixed. For instance, the entries in $\mathbf{m}_{it}$ might include investment, as in Olley and Pakes (1996), or consumption of intermediate inputs (raw materials, energy, etc), as in Levinsohn and Petrin

[1] Of course, it is possible to argue that price-cost markups are responsible for any discrepancy that is observed between the time-invariant elasticity parameters and their corresponding input expenditure ratios. This possibility is not considered here, although the interested reader may consult Kilinc (2014).

[2] It is assumed, however, that $\mathbf{z}_{it}$, $\mathbf{x}_{it}$, and $\mathbf{w}_{it}$ do not have any overlapping elements.

(2003). The intuition that underlies this approach is rather straightforward: for a given quantity of quasi-fixed inputs, more productive firms tend to be more profitable, and hence they have more reason to invest in the expansion of their capital stock. Similarly, firms who are more productive are *ceteris paribus* able to produce more output; as a result, they tend to require greater quantities of raw materials and energy. In what follows, we adhere to the same procedure that is delineated in Wooldridge (2009), who expresses the vector of proxy variables as a function of the period-t state variables and the unobservable productivity term $\mathbf{m}_{it} = m(\mathbf{x}_{it}, v_{it})$. Moreover, given the assumption that the entries in $\mathbf{m}_{it}$ are strictly increasing in v_{it}, holding $\mathbf{x}_{it}$ fixed, the function $m(\mathbf{x}_{it}, \cdot)$ is invertible. Let the function $g(\mathbf{x}_{it}, \mathbf{m}_{it}) = m^{-1}(\mathbf{x}_{it}, v_{it})$ denote the inverse of $m(\mathbf{x}_{it}, v_{it})$. The unobservable productivity term v_{it} can then be expressed as a function of the vector of quasi-fixed inputs and the vector of proxy variables:

$$v_{it} = g(\mathbf{x}_{it}, \mathbf{m}_{it}), \tag{2.4}$$

where the functional form of $g(\cdot)$ has not yet been specified in a parametric sense.

In addition to (2.4), it is typically assumed in the production function literature that firm productivity follows a first-order Markov process:

$$\begin{aligned} v_{it} &= \mathbb{E}(v_{it}|v_{it-1}) + a_{it} \\ &= f(v_{it-1}) + a_{it}, \end{aligned} \tag{2.5}$$

where a_{it} denotes the unanticipated component of firm i's productivity in period t and once again, the functional form of $f(\cdot)$ is not parametrically-

specified. Therefore, by combining (2.5) and (2.4), it is possible to re-express the production function in (2.3) as:

$$y_{it} = \mathbf{x}_{it} \cdot \beta(\mathbf{z}_{it}) + \mathbf{w}_{it} \cdot \gamma(\mathbf{z}_{it}) + f[g(\mathbf{x}_{it-1}, \mathbf{m}_{it-1})] + a_{it} + e_{it}. \tag{2.6}$$

Note that (2.6) is the varying coefficient counterpart of equation (2.9) in Wooldridge (2009).

2.2.1 Estimation strategy

Even though (2.6) bears a close resemblance to Robinson's (1988) partially linear model, it cannot be estimated using the standard semiparametric regression approach because 1) the elasticity coefficients β and γ are not fixed parameters, and are instead functions of the vector $\mathbf{z}_{it}$, and 2) $\mathbf{w}_{it}$ is not independent of a_{it}. Nevertheless, as a first step in estimating (2.6), one can follow a similar approach to that of Robinson (1988) and take the conditional expectation of y_{it}:

$$\begin{aligned}\mathbb{E}(y_{it}|\mathbf{x}_{it-1}, \mathbf{m}_{it-1}, \mathbf{z}_{it}) &= \mathbb{E}(\mathbf{x}_{it}|\mathbf{x}_{it-1}, \mathbf{m}_{it-1}, \mathbf{z}_{it}) \cdot \beta(\mathbf{z}_{it}) \\ &\quad + \mathbb{E}(\mathbf{w}_{it}|\mathbf{x}_{it-1}, \mathbf{m}_{it-1}, \mathbf{z}_{it}) \cdot \gamma(\mathbf{z}_{it}) + f[g(\mathbf{x}_{it-1}, \mathbf{m}_{it-1})],\end{aligned} \tag{2.7}$$

where (2.7) makes use of the fact that $\mathbb{E}(a_{it} + e_{it}|\mathbf{x}_{it-1}, \mathbf{m}_{it-1}, \mathbf{z}_{it}) = 0$. Subtracting (2.7) from (2.6) then yields the following:

$$\tilde{y}_{it} = \tilde{\mathbf{x}}_{it} \cdot \beta(\mathbf{z}_{it}) + \tilde{\mathbf{w}}_{it} \cdot \gamma(\mathbf{z}_{it}) + a_{it} + e_{it}, \tag{2.8}$$

where (2.8) uses the notational convention $\tilde{y}_{it} = y_{it} - \mathbb{E}\left(y_{it}|\mathbf{x}_{it-1}, \mathbf{m}_{it-1}, \mathbf{z}_{it}\right)$, $\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \mathbb{E}\left(\mathbf{x}_{it}|\mathbf{x}_{it-1}, \mathbf{m}_{it-1}, \mathbf{z}_{it}\right)$, and $\tilde{\mathbf{w}}_{it} = \mathbf{w}_{it} - \mathbb{E}\left(\mathbf{w}_{it}|\mathbf{x}_{it-1}, \mathbf{m}_{it-1}, \mathbf{z}_{it}\right)$. In practice, $\tilde{y}_{it}$, $\tilde{\mathbf{x}}_{it}$, and $\tilde{\mathbf{w}}_{it}$ can respectively be estimated by taking the residuals from nonparametric regressions of y_{it}, $\mathbf{x}_{it}$, and $\mathbf{w}_{it}$ on $\mathbf{x}_{it-1}$, $\mathbf{m}_{it-1}$, and $\mathbf{z}_{it}$. In this chapter, these regressions are carried out using the Nadaraya-Watson local constant method of Racine and Li (2004) in Hayfield and Racine's (2008) 'np' package in the R statistical computing environment.

Given that it is assumed in this chapter that the firm makes its variable input decision *after* observing the unanticipated productivity shock a_{it} at the beginning of period t, $\tilde{\mathbf{w}}_{it}$ in equation (2.8) is endogenous. On an intuitive level, the dependence of $\mathbf{w}_{it}$ on a_{it} stems from the fact that the latter influences the marginal product of the former which, holding the output price fixed, has implications for the firm's optimal choice of variable input quantities. Fortunately, finding a suitable instrument does not present much of a challenge since $\tilde{\mathbf{w}}_{it-1} = \mathbf{w}_{it-1} - \mathbb{E}\left(\mathbf{w}_{it-1}|\mathbf{x}_{it-1}, \mathbf{m}_{it-1}\right)$ is in most instances strongly correlated with $\tilde{\mathbf{w}}_{it}$, and it is necessarily independent of a_{it}.[3] At this juncture, the objective is to make use of the orthogonality assumptions $\mathbb{E}\left(a_{it} + e_{it}|\tilde{\mathbf{x}}_{it}\right) = 0$ and $\mathbb{E}\left(a_{it} + e_{it}|\tilde{\mathbf{w}}_{it-1}\right) = 0$ in order to identify the functional elasticity coefficients $\beta\left(\mathbf{z}_{it}\right)$ and $\gamma\left(\mathbf{z}_{it}\right)$. Adopting the shorthand notation $\boldsymbol{\delta}\left(\mathbf{z}_{it}\right) = \left[\beta\left(\mathbf{z}_{it}\right)' \quad \gamma\left(\mathbf{z}_{it}\right)'\right]'$ for the vector of varying elasticity coefficients, $\tilde{\mathbf{v}}_{it} = [\tilde{\mathbf{x}}_{it} \quad \tilde{\mathbf{w}}_{it}]$ for the vector of endogenous regressors, and $\overset{\circ}{\tilde{\mathbf{v}}}_{it} = [\tilde{\mathbf{x}}_{it} \quad \tilde{\mathbf{w}}_{it-1}]$ for the vector of instruments, the sample moment condition that corresponds

[3] Li and Stengos (1996) propose a similar solution to the problem of endogeneity in partial linear models.

to $\mathbb{E}\left(a_{it} + e_{it}|\tilde{\mathbf{x}}_{it}\right) = 0$ and $\mathbb{E}\left(a_{it} + e_{it}|\tilde{\mathbf{w}}_{it-1}\right) = 0$ can be written as:

$$\frac{1}{NT}\sum_{i}\sum_{t}\overset{\circ}{\tilde{\mathbf{v}}}'_{it}\left(\tilde{y}_{it} - \tilde{\mathbf{v}}_{it}\boldsymbol{\delta}\left(\mathbf{z}_{it}\right)\right)\boldsymbol{\omega}_{it} = 0, \tag{2.9}$$

where $\boldsymbol{\omega}_{it}$ is a local weighting function.

The exact functional form of $\boldsymbol{\omega}_{it}$ will depend on the variables that are included in $\mathbf{z}_{it}$, although it can be expressed using Li and Racine's (2010) general notation of a product kernel for mixed data types:

$$\boldsymbol{\omega}_{it} = \prod_{j=1}^{q}\omega_{j,it}^{c}\prod_{k=1}^{r}\omega_{k,it}^{u}\prod_{l=1}^{s}\omega_{l,it}^{o}, \tag{2.10}$$

where the ω_{it}^{c}'s, ω_{it}^{u}'s, and ω_{it}^{o}'s respectively weight observations in the sample based on their values of the $q + r + s$ continuous, unordered discrete, and ordered discrete variables that influence the coefficients β and γ. For a given continuous variable z^{c} in the vector $\mathbf{z}$, the kernel weighting function is defined as $\omega_{it}^{c} = \frac{1}{h}\cdot k\left(\frac{z_{it}^{c} - z^{c}}{h}\right)$, where h is a smoothing parameter whose value is chosen according to some sort of selection criterion, and the bounded, symmetric kernel $k\left(\cdot\right)$ satisfies $\int k\left(\varphi\right)d\varphi = 1$, $\int \varphi^{2}k\left(\varphi\right)d\varphi > 0$, and $\int k^{2}\left(\varphi\right)d\varphi > 0$. In most applications, a second-order Gaussian kernel is used for $k\left(\cdot\right)$. Next, the weighting function ω_{it}^{u} that corresponds to a given unordered discrete variable z^{u} in $\mathbf{z}$ is defined according to:

$$\omega_{it}^{u} = \begin{cases} 1 & \text{if } \; z_{it}^{u} = z^{u} \\ \mu & \text{if } \; z_{it}^{u} \neq z^{u}, \end{cases} \tag{2.11}$$

where $\mu \in [0,1]$ is once again a smoothing parameter whose optimal value can be determined using a variety of methods (more on this later). Finally, the function that weights observations in the sample in relation to the ordered discrete variable z^o in $\mathbf{z}$ is expressed as

$$\omega_{it}^o = \begin{cases} 1 & \text{if } \; z_{it}^o = z^o \\ \lambda^{|z_{it}^o - z^o|} & \text{if } \; z_{it}^o \neq z^o \end{cases} \tag{2.12}$$

for some $\lambda \in [0,1]$. Note that (2.11) and (2.12) bear a close resemblance to one another; their sole distinguishing feature, aside from the fact that their respective smoothing parameters μ and λ will generally be assigned different values, is that the former attaches no meaning to the "distance"between z_{it}^u and z^u, whereas the latter treats $|z_{it}^o - z^o|$ as significant.

For a given product kernel $\boldsymbol{\omega}_{it}$ with smoothing parameters $(h'_{q\times 1}, \mu'_{r\times 1}, \lambda'_{s\times 1})$, simplification of the sample moment condition in (2.9) yields

$$\hat{\boldsymbol{\delta}}\left(\mathbf{z}_{it}\right) = \left[\hat{\beta}\left(\mathbf{z}_{it}\right)' \quad \hat{\gamma}\left(\mathbf{z}_{it}\right)'\right]' = \left[\sum_i \sum_t \mathring{\tilde{\mathbf{v}}}'_{it} \tilde{\mathbf{v}}_{it} \boldsymbol{\omega}_{it}\right]^{-1} \left[\sum_i \sum_t \mathring{\tilde{\mathbf{v}}}'_{it} \tilde{y}_{it} \boldsymbol{\omega}_{it}\right]. \tag{2.13}$$

Equation (2.13) is effectively an application of the smooth coefficient generalized method of moments framework proposed by Tran and Tsionas (2010) and the class of varying coefficient models with mixed data types proposed by Li and Racine (2010). Estimation is carried out using the kernel summation function in Hayfield and Racine's (2008) 'np' package in the R statistical computing environment.

2.2.2 Bandwidth selection

While providing empirical applications of their smooth coefficient GMM estimator, Tran and Tsionas (2010) rely on a "rule-of-thumb" criterion for bandwidth selection. Although this approach is not entirely without theoretical justification[4], it has several deficiencies that make it unsuitable for the estimation of (2.13). Most importantly, the rule-of-thumb criterion is only applicable to functions of a continuous variable; if, on the other hand, estimation of the coefficients β and γ involves a weighting function such as (2.10), then an alternative bandwidth selection procedure must be followed.

As a second option, regardless of whether the vector $\mathbf{z}_{it}$ consists of continuous or discrete data (or both), the optimal bandwidth vector $(h^{*\prime}_{q\times 1}, \mu^{*\prime}_{r\times 1}, \lambda^{*\prime}_{s\times 1})$ can be selected using data-driven methods such as least-squares cross-validation:

$$\left(h^{*\prime}_{q\times 1}, \mu^{*\prime}_{r\times 1}, \lambda^{*\prime}_{s\times 1}\right) = \underset{h,\mu,\lambda}{\arg\min} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\tilde{y}_{it} - \tilde{\mathbf{v}}_{it} \hat{\boldsymbol{\delta}}_{-it}\left(\mathbf{z}_{it}\right)\right]^2, \tag{2.14}$$

where $\hat{\boldsymbol{\delta}}_{-it}$ is a "leave-one-out"estimator that is computed using a variation of (2.13) in which observation i, t has been excluded from the sample, and evaluated at $\mathbf{z}_{it}$.[5] The least-squares cross-validation procedure is analogous to minimizing the integrated mean squared error of the estimated smooth coefficient regression function, and a detailed explanation of its properties can be found in Li and Racine (2004, 2007). The constrained minimization routine that is used to compute the optimal bandwidth vector in (2.14) is implemented using the 'nlminb' command in the R statistical computing environment.

[4] See, for instance, Silverman (1986).

[5] For notational simplicity, (2.14) applies to the special case of a balanced panel dataset.

2.2.3 Plant productivity

The discussion up to this point has centered on the identification and estimation of the elasticity coefficients in (2.3); however, it remains to be seen what the implications are for the analysis of firm-level productivity. Consider once again the Cobb-Douglas model in (2.3). The unobservable plant productivity term, v_{it}, can be expressed as:

$$v_{it} = y_{it} - \mathbf{x}_{it} \cdot \beta(\mathbf{z}_{it}) - \mathbf{w}_{it} \cdot \gamma(\mathbf{z}_{it}) - e_{it}, \tag{2.15}$$

where the reader will recall that e_{it} is a random disturbance term whose mean conditional on current and past inputs is equal to zero. If one follows the procedure that is articulated in Section 2.2, obtaining estimates of the elasticity vectors $\beta(\mathbf{z}_{it})$ and $\gamma(\mathbf{z}_{it})$ is straightforward. Let $\hat{r}_{it} = y_{it} - \mathbf{x}_{it} \cdot \hat{\beta}(\mathbf{z}_{it}) - \mathbf{w}_{it} \cdot \hat{\gamma}(\mathbf{z}_{it})$. One can then write $\hat{r}_{it} = v_{it} + e_{it}$. Using the proxy expression $v_{it} = g(\mathbf{x}_{it}, \mathbf{m}_{it})$ from (2.4), where the precise functional form of $g(\cdot)$ is unknown, and the conditional expectation $\mathbb{E}(e_{it}|\mathbf{x}_{it}, \mathbf{m}_{it}) = 0$, the nonparametric regression equation below can be used to estimate plant-level productivity:

$$\hat{r}_{it} = g(\mathbf{x}_{it}, \mathbf{m}_{it}) + e_{it}. \tag{2.16}$$

Equation (2.16) can be estimated using, for example, the nonparametric local constant (i.e. Nadaraya-Watson) regression method.

2.3 Empirical example

The unbalanced panel dataset that underlies the empirical analysis that follows consists of 54,545 observations on Colombian manufacturing plants from 22 different industries[6] and is obtained from an annual census that was conducted during the period 1981-1991. The advantage of using this plant-level data from Colombia is that it has been used in a number of other studies for the estimation of firm-level production functions, and hence the results that are reported in Section 2.3.2 can be readily contrasted with what is found elsewhere in the literature.[7] To begin, the value-added term y_{it} on the left-hand side of (2.6) is computed as the difference between the gross value of firm i's output and its intermediate consumption in period t. The latter is defined as the sum of energy purchased, raw materials consumed, and miscellaneous industrial expenditures incurred.[8]

Next, firm i's capital stock in period t is assumed to be the sole quasi-fixed input x_{it} in the model. The capital stock variable comprises land, buildings/structures, machinery/equipment, transportation equipment, and office equipment, whose respective values are computed using the perpetual inventory method and annual 3-digit industry-level depreciaton data found in Pombo (1999). The assumption that physical capital is quasi-fixed stems from the fact that the construction of new buildings and structures, as well as the

[6]Industries are differentiated based on the 3-digit International Standard Industrial Classification (ISIC) system. The 22 3-digit industries can be further subdivided into 79 unique 4-digit ISIC codes.

[7]See for instance Roberts and Tybout (1997), Fernandes (2007), and Gandhi et al. (2016).

[8]The miscellaneous industrial expenditures include, for instance, purchases of fuels and lubricants, purchases of accessories and replacement parts of less than one year duration, and payments to third parties for repairs and maintenance.

purchase and installation of new machinery and equipment, tend to require a rather significant time commitment. As a result, firms are unlikely to be able to make these types of input decisions with complete flexibility in the short term. Moving on, the variable input vector $\mathbf{w}_{it} = [w_{s,it} \;\; w_{u,it}]$ in (2.6) has two entries that are respectively defined as the total number of skilled and unskilled workers employed by the firm in a particular year. The census of Colombian manufacturers specifies five categories of employees: managers, skilled workers, technicians, unskilled workers, and apprentices. In the data that is used to estimate the smooth coefficient production function in (2.6), managers, skilled workers, and technicians are classified as "skilled", whereas unskilled workers and apprentices are classified as "unskilled." Next, following the same blueprint as Levinsohn and Petrin (2003), firm i's period-t consumption of energy serves as the scalar proxy variable m_{it} that is used to generate a function controlling for unobservable productivity v_{it} in (2.1). To restate what was mentioned earlier in Section 2.2, the assumption that underlies this choice of a proxy variable is that more productive firms, when their capital stock is held fixed, are able to manufacture more output and hence, they need to consume more intermediate inputs such as energy. The energy consumption variable is measured in thousands of Colombian pesos. As a final note, it should be pointed out that the value-added (y_{it}), capital stock (x_{it}), and energy consumption (m_{it}) variables are all deflated by an annual 3-digit industry-level price index prior to taking their logs in order to normalize input and output quantities across time periods. In the census data, both the nominal and the real values of a firm's total output are reported annually, and hence the price index is obtained by taking the ratio thereof. For additional details vis-a-vis

the panel dataset of Colombian manufacturers and the multi-year census from which it originates, the interested reader is encouraged to consult Roberts (1996).

2.3.1 Functional specification of the elasticity coefficients

Given that the dataset described above comprises observations from a number of different industries, not much information will be gleaned from the estimation of an aggregate production function using the pooled sample of plants whose final goods and input requirements are likely quite heterogeneous. The convention that is followed in much of the literature is to estimate separate production functions using industry-level subsamples of a given dataset; however, one of the disadvantages of this approach is a reduction in sample size that might lead to less precise estimates of the various industries' elasticity coefficients. Alternatively, one may choose to keep the pooled sample of firms intact and use kernel methods to weight individual observations based on their industry identifier. Similarly, in light of the fact that the panel of Colombian manufacturing plants spans an 11-year period, it might be wise to loosen the assumption that production technology is static and instead allow the elasticity coefficients β and γ to vary over time. Hence, the model that is estimated here is a varying coefficient production function for plant i in industry j and period t:

$$y_{it} = x_{it} \cdot \beta_{jt} + w_{u,it} \cdot \gamma_{u,jt} + w_{s,it} \cdot \gamma_{s,jt} + f\left[g_j\left(x_{it-1}, m_{it-1}\right)\right] + a_{it} + e_{it}, \quad (2.17)$$

where, to reiterate, x_{it}, $w_{u,it}$, and $w_{s,it}$ denote the natural log of plant i's period-t capital stock and quantity of unskilled (u) and skilled (s) workers, respectively. In terms of the notation from Section 2.2, $\mathbf{z}_{it}$ would be a two-dimensional vector that comprises the industry and time indicators, which are treated as unordered and ordered discrete variables, respectively. Hence, the function $\boldsymbol{\omega}_{it}$ that weights observations in the sample according to $\{j, t\}$ is defined as the product of the kernels in (2.11) and (2.12). Note that the proxy function $g_j(\cdot)$ controlling for unobservable productivity is now assumed to be industry-specific, which is why it is indexed by the subscript j along with the time-varying elasticity coefficients β_{jt}, $\gamma_{u,jt}$, and $\gamma_{s,jt}$. As a means of reducing the variance of the estimate of β_{jt} arising from a very high degree of correlation between x_{it} and x_{it-1}, the firm's stock of machinery and equipment in period $t-1$ rather than it's entire capital stock is denoted by x_{it-1}. In what follows, the empirical approach that is outlined in Section 2.2.1 is applied to the estimation of (2.17), and the various outcomes of this exercise are discussed.

2.3.2 Elasticity coefficient estimates

Although the census of Colombian manufacturers collects data on plants who operate in 22 different industries, much of the production function literature has tended to focus on four industries that are deemed particularly important for South American economies. For instance, Levinsohn and Petrin (2003), Ackerberg et al. (2006), and Gandhi et al. (2016), all center their analyses on food processing (ISIC=311), textiles (ISIC=321), finished wood products (ISIC=331), and fabricated metal (ISIC=381).[9] In accordance with this con-

[9]The apparel industry (ISIC=322) is also considered in Gandhi et al.'s (2016) analysis of Colombian manufacturing plants but is omitted from the analyses of Levinsohn and Petrin

vention, the discussion that follows revolves around these four key industries.

Figures 2.1 and 2.2 comprise plots of the elasticity coefficient estimates $\hat{\beta}_{jt}$, $\hat{\gamma}_{u,jt}$, and $\hat{\gamma}_{s,jt}$, as well as their bootstrapped 90 percent confidence intervals, over time. For comparative purposes, these plots also include superimposed estimates of β, γ_u, and γ_s, obtained using the method of Li and Stengos (1996), under the scenario where these parameters are treated as fixed. The first four plots that appear in figure 2.1 suggest that Colombian food processors' production technology, particularly in terms of physical capital and unskilled labour, underwent some significant changes between 1982 and 1991. The time-varying estimates of β_{jt}, which denotes the elasticity of output with respect to physical capital, are steadily on the rise throughout the decade, starting at 0.1258 in 1982 and eventually reaching a peak of 0.2210 in 1991. Meanwhile, there is a very noticeable downward trend in the estimated elasticity of output with respect to unskilled labour; specifically, over a ten-year period, $\hat{\gamma}_{u,jt}$ falls from a high of 0.5629 in 1982 all the way down to 0.2945 in 1991. The year-by-year estimates of $\gamma_{s,jt}$ do not exhibit the same dramatic movements as $\hat{\gamma}_{u,jt}$, and instead are characterized by a comparatively subtle U-shaped pattern during the 1982-1991 period. Altogether, over the course of a decade, there is a substantial decline in the estimated returns to scale, which start out at 1.08 and end up at just 0.885 by the early 1990s.

Coefficient estimates for the textile industry are presented in the last four plots of figure 2.1, while the $\hat{\beta}_{jt}$, $\hat{\gamma}_{u,jt}$, and $\hat{\gamma}_{s,jt}$ for the wood products and fabricated metal industries are plotted in the top-half and bottom-half portions of figure 2.2, respectively. Overall, the changes that are observed in the textile

(2003) and Ackerberg et al. (2006), who use data from the Chilean manufacturing sector.

industry's production function over a ten-year period are qualitatively similar to what was noted for food processors. Specifically, the estimated elasticity of output with respect to capital is higher in the early 1990s than it was at the outset in 1982. Over the ten-year period that is covered in the sample, $\hat{\beta}_{jt}$ increases by about 0.07, while the change in $\hat{\gamma}_{s,jt}$ is comparatively small. Meanwhile, in the textile, wood products, and fabricated metal industries, there appears to be a rather substantial downward trend in the contribution of unskilled labour to value-added output, as attested to by the respective decreases in $\hat{\gamma}_{u,jt}$ of 0.218, 0.237, and 0.188 between 1982 and 1991. Returns to scale estimates fall by a similar magnitude. It is interesting to note that these latter results closely mimic what was reported for food processing plants.

2.4 Monte Carlo experiment

In this subsection, a series of Monte Carlo experiments with M=100 draws in each case are undertaken as a means of evaluating the extent to which the proposed semiparametric framework is able to capture different levels of variation in parameters of the type that appear in (2.17), which we continue to denote by β_{jt}, $\gamma_{u,jt}$, and $\gamma_{s,jt}$. Sample sizes $n = 500$, 1000, 2500, 5000, and 10000 are considered for a version of the model shown in (2.6) where x_{it}, x_{it-1}, and m_{it-1} are all univariate, and $\mathbf{w}_{it} = [w_{u,it} \;\; w_{s,it}]$, $\mathbf{w}_{it-1} = [w_{u,it-1} \;\; w_{s,it-1}]$ are both two-dimensional. The five variables that are assumed to be exogenous, namely x_{it}, x_{it-1}, m_{it-1}, $w_{u,it-1}$, and $w_{s,it-1}$, are drawn from a multivariate Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, where $\mathbf{\Sigma}_{i,i} = 1$ for $i = 1, ..., 5$ and $\mathbf{\Sigma}_{i,j} = 0.4$ for all $i \neq j$. The random disturbance term is distributed $e_{it} \sim \mathcal{N}(0, .2)$, while the endogenous $w_{u,it}$ and $w_{s,it}$ are

respectively given by $w_{u,it} = 0.5 + 0.75w_{u,it-1} + 0.5e_{it}$ and $w_{s,it} = 0.85w_{u,it-1} + 0.6e_{it}$. The analogs of the industry and time variables are drawn from discrete uniform distributions with possible values $j \in \{1, 2, 3\}$ and $t \in \{0, 1, ..., 10\}$, respectively. Three distinct scenarios relating to the functional form of β_{jt}, $\gamma_{u,jt}$, and $\gamma_{s,jt}$ are considered:

1. $\beta_{jt} = \frac{j}{2} \cdot \sin\left(\frac{\pi t}{5}\right), \quad \gamma_{u,jt} = \frac{j}{2} \cdot \left(-\cos\left(\frac{\pi t}{10}\right)\right), \quad \gamma_{s,jt} = \frac{j}{2} \cdot \cos\left(\frac{\pi t}{5}\right)$

2. $\beta_{jt} = \frac{j}{2} \cdot \left(\frac{t}{10}\right)^3, \quad \gamma_{u,jt} = \frac{j}{2} \cdot \left(1 - \frac{t}{10}\right), \quad \gamma_{s,jt} = \frac{j}{2} \cdot \sin\left(\frac{\pi t}{10}\right)$

3. $\beta_{jt} = \frac{j}{2} \cdot 0.35, \quad \gamma_{u,jt} = \frac{j}{2} \cdot \frac{t}{10}, \quad \gamma_{s,jt} = \frac{j}{2} \cdot 0.65$

Note that in this example, it is assumed that the industry identifier j affects the magnitude but not the curvature of the elasticity functions. Under the first two scenarios that are outlined above, there is a considerable amount of time variation in β_{jt}, $\gamma_{u,jt}$, and $\gamma_{s,jt}$ and furthermore, the functions that underlie these coefficients exhibit contrasting degrees of curvature. Meanwhile, the third functional form specification covers the special case where two of the three elasticity coefficients do not vary from one time period to another, so that it can be ascertained that the kernel weighting framework proposed in Section 2.2.1 does not yield time-varying coefficient estimates when they are not in fact found in the underlying data generating process. Given that it is somewhat peripheral to the core analysis being undertaken here, the unknown composite function $f\left[g_j\left(x_{it-1}, m_{it-1}\right)\right]$ from (2.6) is straightforwardly specified in all three cases as $f\left[g_j\left(x_{it-1}, m_{it-1}\right)\right] = j \cdot 0.2 \cdot (x_{it-1} - m_{it-1})$.

Figure 2.3 illustrates the respective outcomes of the Monte Carlo experiments with sample size $n = 2500$. The median coefficient estimates at each $t \in \{0, 1, ..., 10\}$, holding $j = 2$ fixed, are represented by small red x's, while

the shaded vertical line segments at each t reflect the 90 percent confidence intervals for β_{jt}, $\gamma_{u,jt}$, and $\gamma_{s,jt}$. The estimation procedure outlined in Section 2.2.1 appears to perform rather well in finite sample environments where the data-generating process is known to the researcher. Table 2.2 shows that under all 3 model specifications, the average mean squared error (AMSE) of the different coefficient estimates shrinks toward zero as the sample size grows incrementally from $n = 500$ to $n = 10000$. Unsurprisingly, the AMSE tends to be larger for the coefficients that exhibit the most variation, although even under the first scenario where β_{jt}, $\gamma_{u,jt}$, and $\gamma_{s,jt}$ are all periodic functions of t, the semiparametric method yields a good fit of the model.

2.5 Conclusion

This chapter has delineated a semiparametric method of estimating a firm-level Cobb-Douglas production function with varying elasticity coefficients that are determined by a combination of continuous and discrete predictors. The varying coefficient approach is more consistent with the theoretical relationship that ought to exist between factor elasticities and their corresponding plant-level input expenditure ratios. Using a dataset from Colombia that has proven popular elsewhere in the literature, it has been shown that production technology can exhibit considerable variation over time. In particular, the elasticity of output with respect to unskilled labour in four key manufacturing industries appears to have been in decline during the ten-year period between 1982 and 1991. This gave rise to a rather sharp drop in estimated returns to scale. Altogether, the semiparametric estimator performs well in a Monte Carlo setting; when the true data generating process is known, it yields fairly precise

estimates of varying coefficients under a number of different specifications of the model. Implications for the measurement of firm productivity have been briefly considered, although this portion of the analysis can likely be expanded upon in future research.

References

Ackerberg, D., K. Caves, and G. Frazer (2006): "Structural identification of production functions," Mpra paper, University Library of Munich, Germany.

Arellano, M. and S. Bond (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, 58, 277–97.

Blundell, R. and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of Econometrics*, 87, 115–143.

——— (2000): "GMM estimation with persistent panel data: an application to production functions," *Econometric Reviews*, 19, 321–340.

Cai, Z. and Q. Li (2008): "Nonparametric estimation of varying coefficient dynamic panel data models," *Econometric Theory*, 24, 1321–1342.

Chen, R. and R. Tsay (1993): "Functional coefficient autoregressive models," *Journal of the American Statistical Association*, 88, 298–308.

Cobb, C. W. and P. H. Douglas (1928): "A theory of production," *American Economic Review*, 18, 139–172.

Fernandes, A. (2007): "Trade policy, trade volumes and plant-level productivity in Colombian manufacturing industries," *Journal of International Economics*, 71, 52–71.

Gandhi, A., S. Navarro, and D. Rivers (2016): "On the identification of production functions: How heterogeneous is productivity?" Manuscript.

Griliches, Z. and J. Mairesse (1995): "Production functions: The search for identification," NBER Working Papers 5067, National Bureau of Economic Research, Inc.

Hastie, T. and R. Tibhirani (1993): "Varying coefficient models," *Journal of the Royal Statistical Society Series B*, 55, 757–796.

Hayfield, T. and J. S. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27.

Huang, J. Z., C. O. Wu, and L. Zhou (2004): "Polynomial spline estimation and inference for varying coefficient models with longitudinal data," *Statistica Sinica*, 14, 763–788.

Kilinc, U. (2014): "Estimating entrants' productivity when prices are unobserved," *Economic Modelling*, 38, 640–647.

Lee, Y. K., E. Mammen, and B. U. Park (2012): "Flexible generalized varying coefficient regression models," *Annals of Statistics*, 40, 1906–1933.

Levinsohn, J. and A. Petrin (2003): "Estimating production functions using inputs to control for unobservables," *Review of Economic Studies*, 70, 317–341.

Li, Q., C. J. Huang, D. Li, and T. Fu (2002): "Semiparametric smooth coefficient models," *Journal of Business and Economic Statistics*, 20, 412–422.

Li, Q. and J. Racine (2010): "Smooth varying-coefficient estimation and

inference for qualitative and quantitative data," *Econometric Theory*, 26, 1607–1637.

Li, Q. and J. S. Racine (2004): "Cross-validated local linear nonparametric regression," *Statistica Sinica*, 14, 485–512.

——— (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Li, Q. and T. Stengos (1996): "Semiparametric estimation of partially linear panel data models," *Journal of Econometrics*, 71, 389–397.

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

Olley, S. G. and A. Pakes (1996): "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 64, 1263–97.

Park, B. U., E. Mammen, Y. K. Lee, and E. R. Lee (2015): "Varying coefficient regression models: A review and new developments," *International Statistical Review*, 83, 36–64.

Pombo, C. (1999): "Productividad industrial en Colombia: Una aplicacion de numeros indices," *Revista de Economia de la Universidad del Rosario*, 107–139.

Racine, J. S. and Q. Li (2004): "Nonparametric estimation of regression functions with both categorical and continuous data," *Journal of Econometrics*, 119, 99–130.

Roberts, M. (1996): "Colombia 1977-1985: Producer turnover, margins, and trade exposure," in *Industrial Evolution in Developing Countries: Micro Patterns of Turnover, Productivity, and Market Structure*, ed. by J. M. Roberts and R. J. Tybout, Oxford University Press.

Roberts, M. and J. Tybout (1997): "The decision to export in Colombia: An empirical model of entry with sunk costs," *American Economic Review*, 87, 545–64.

Robinson, P. (1988): "Root-N-consistent semiparametric regression," *Econometrica*, 56, 931–54.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Tran, K. and E. Tsionas (2010): "Local GMM estimation of semiparametric panel data with smooth coefficient models," *Econometric Reviews*, 29, 39–61.

Wooldridge, J. (2009): "On estimating firm-level production functions using proxy variables to control for unobservables," *Economics Letters*, 104, 112–114.

Appendix

	Food processing	Textiles	Wood products	Fabricated metal
β				
VC min	0.126	0.101	0.062	0.081
VC max	0.223	0.181	0.104	0.112
CC	0.189	0.141	0.087	0.084
γ_u				
VC min	0.294	0.344	0.430	0.344
VC max	0.563	0.566	0.667	0.532
CC	0.400	0.402	0.525	0.400
γ_s				
VC min	0.342	0.444	0.346	0.521
VC max	0.391	0.504	0.510	0.552
CC	0.364	0.477	0.395	0.529
RTS				
VC min	0.885	0.979	1.030	0.963
VC max	1.080	1.126	1.112	1.165
CC	0.953	1.021	1.007	1.013

Table 2.1: Summary of varying coefficient (VC) and constant coefficient (CC) elasticity and returns to scale estimates.

		$n = 500$	$n = 1000$	$n = 2500$	$n = 5000$	$n = 10000$
1.	$\beta_{jt} = \frac{j}{2} \cdot \sin\left(\frac{\pi t}{5}\right)$	.03117	.01374	.00490	.00193	.00128
	$\gamma_{u,jt} = \frac{j}{2} \cdot \left(-\cos\left(\frac{\pi t}{10}\right)\right)$	.03477	.01706	.00644	.00244	.00200
	$\gamma_{s,jt} = \frac{j}{2} \cdot \cos\left(\frac{\pi t}{5}\right)$	.04066	.01669	.00605	.00201	.00162
2.	$\beta_{jt} = \frac{j}{2} \cdot \left(\frac{t}{10}\right)^3$	.00907	.00394	.00160	.00065	.00031
	$\gamma_{u,jt} = \frac{j}{2} \cdot \left(1 - \frac{t}{10}\right)$	.01263	.00549	.00217	.00090	.00061
	$\gamma_{s,jt} = \frac{j}{2} \cdot \sin\left(\frac{\pi t}{10}\right)$	.01513	.00625	.00251	.00100	.00048
3.	$\beta_{jt} = \frac{j}{2} \cdot 0.35$	.00300	.00156	.00059	.00028	.00016
	$\gamma_{u,jt} = \frac{j}{2} \cdot \frac{t}{10}$	.01169	.00630	.00229	.00096	.00050
	$\gamma_{s,jt} = \frac{j}{2} \cdot 0.65$	.00648	.00308	.00112	.00036	.00029

Table 2.2: Average mean squared error for 100 Monte Carlo replications.

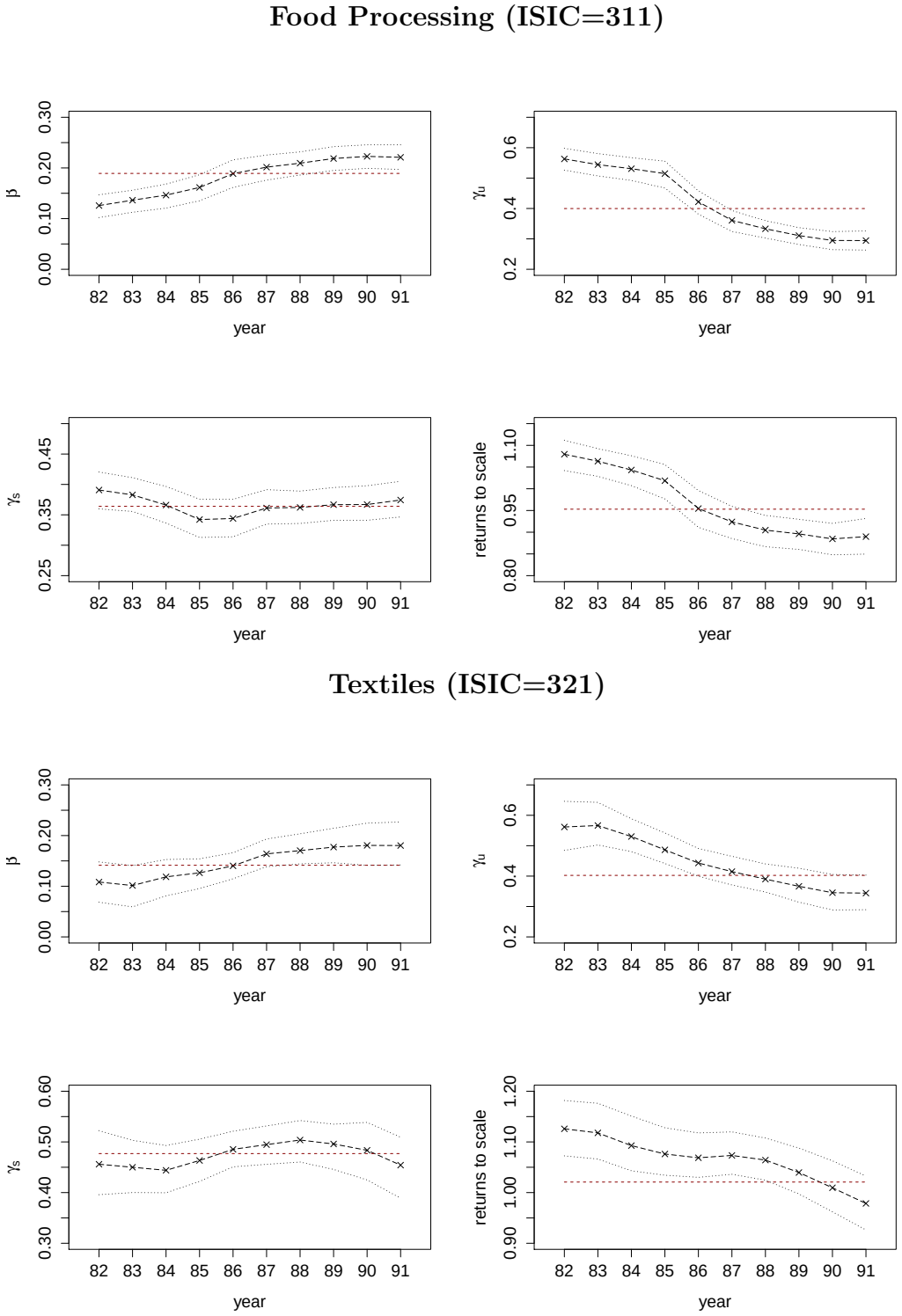

Figure 2.1: Varying coefficient estimates and their bootstrapped 90 percent confidence intervals. Constant coefficient estimates superimposed in red.

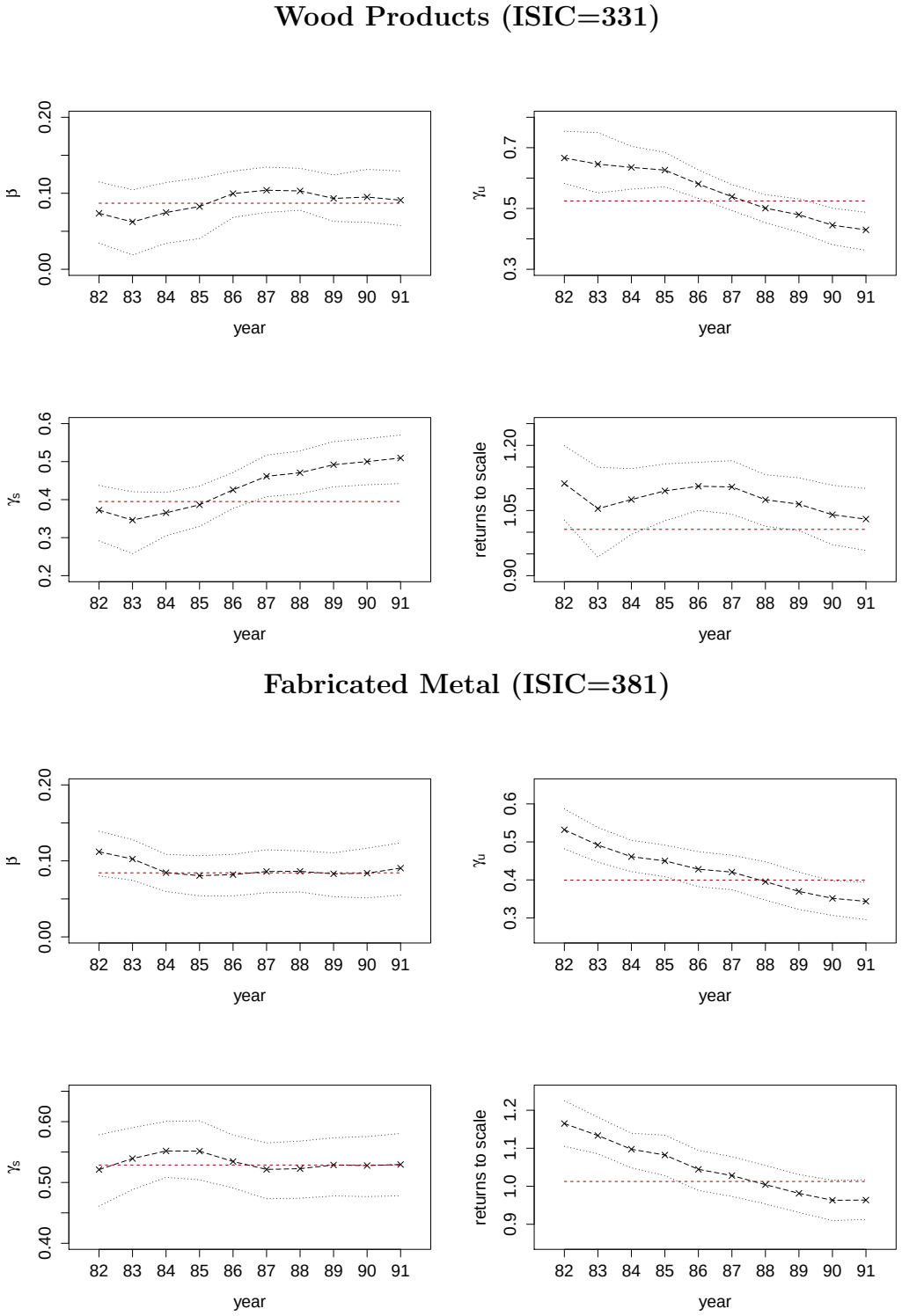

Figure 2.2: Varying coefficient estimates and their bootstrapped 90 percent confidence intervals. Constant coefficient estimates superimposed in red.

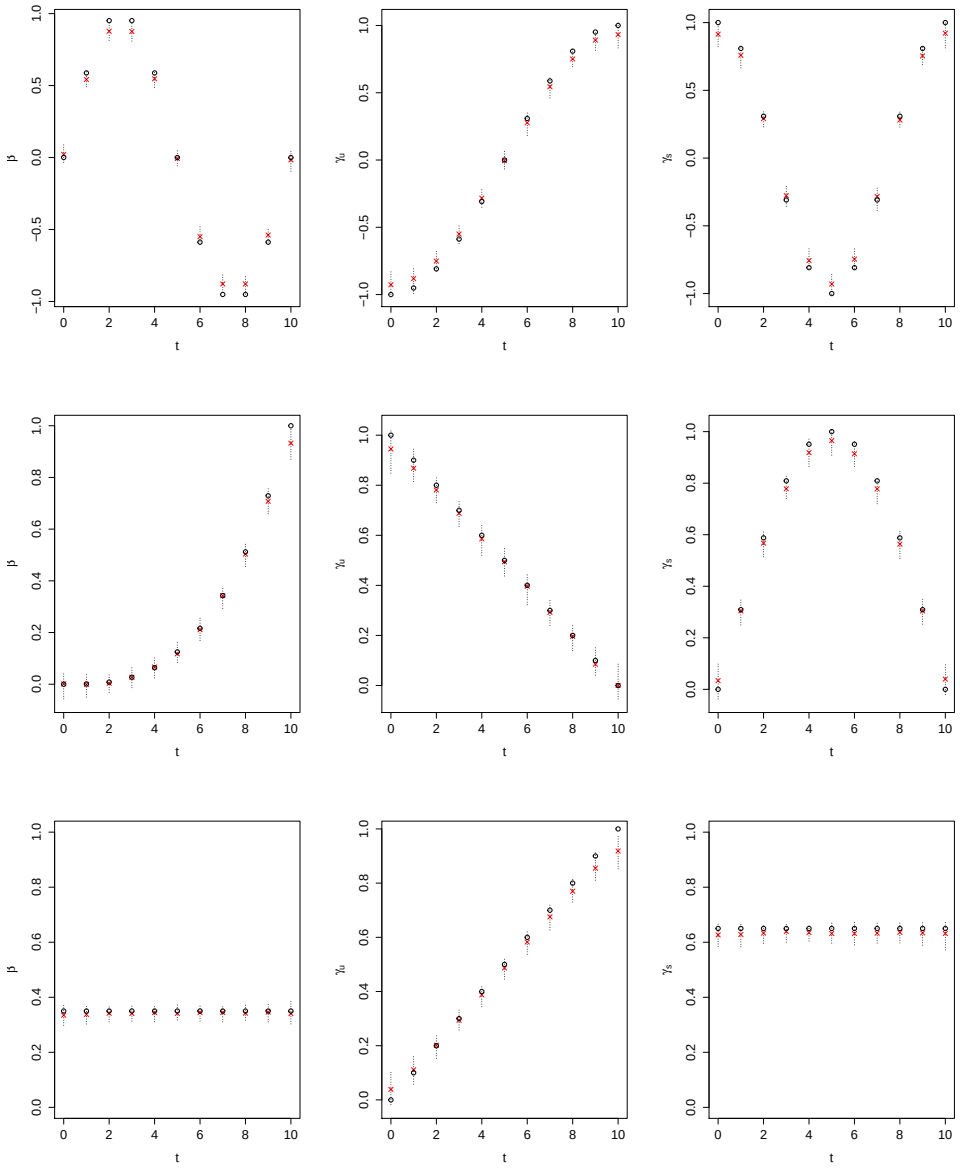

Figure 2.3: Results of Monte Carlo experiment with 100 replications and n=2500: true coefficient values (black circles), mean coefficient estimates (red x's), and 90 percent confidence intervals (shaded lines).

Chapter 3

Robust nonparametric frontier estimation for count data

3.1 Introduction

The construction of a robust methodological framework for the statistical estimation of models of firm-level production remains a central endeavour of modern economic research. While this subfield of econometric analysis is rather diverse, much of it is rooted in an elementary notion from neoclassical production theory, namely the existence of a *frontier function* that conveys how a collection of inputs can be most efficiently combined to produce a particular output. However, an important empirical consideration that is frequently overlooked in the literature is that in some settings, it is more realistic to express a firm's output as a count, rather than a continuous quantity. That is, many production frontier estimators fail to take into explicit consideration the special case of discrete count data, even though it may be the most suitable for

measurement of certain types of industrial output.[1] On the other hand, even if a frontier estimation framework is compatible with a count-valued output, it might be overly-sensitive to outliers, which can arise due to measurement error or some other imperfection in the data collection process. Outliers are a serious concern when applying data envelopment techniques to the extent that they can introduce significant bias into the estimated frontier function. Hence this chapter proposes a statistically-robust nonparametric production frontier estimator for models that involve a count output variable. It improves upon existing robust approaches by making use of a novel data trimming procedure that is based on a combination of k-means and hierarchical clustering.

Contemporary models of firm-level production trace their theoretical roots to the pioneering work of Debreu (1951), Koopmans (1951), and Shephard (1970). In an environment where the vector of inputs x is used to produce output y, one defines the *production set* $\Psi = \{(x, y) : x \text{ can produce } y\}$ which, in turn, gives rise to the notion of a *production frontier* $g(x) = \sup\{y : (x, y) \in \Psi\}$. That is, the function $g(x)$ provides information about the maximum output quantity that can be feasibly produced using the input combination x. Hence, the productive efficiency of a firm, institution, or economic region can be conceptualized in terms of its output shortfall vis-a-vis the frontier. The literature on statistical estimation of frontier functions and productive efficiency is quite vast, and most methodological approaches tend to be categorized as either i) stochastic or ii) deterministic. Given a sample of observed input/output combinations $\{(x_i, y_i)\}_{i=1}^n$, the key distinction that can be drawn between these two classes of models is that the latter requires $(x_i, y_i) \in \Psi$ for $i = 1, ..., n$,

[1] For instance, the output of a mid-cap aircraft manufacturer or the patents awarded to a biotech research firm are clearly counts.

while the former does not. That is, the stochastic frontier framework allows for a certain amount of random noise that may give rise to the inequality $y_i > g(x_i)$ for a limited number of observations. The stochastic frontier literature stems from the early contributions of Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), and is reviewed in detail by Kumbhakar and Lovell (2000) and Parmeter and Kumbhakar (2014). Four papers have focused on models with a count-valued dependent variable. Fe (2013) and Hofler and Scrogin (2008) respectively consider the scenario of an economic bad (good) that ought to be minimized (maximized), although neither framework can be generalized to incorporate both output categories, i.e., both goods and bads. Fe and Hofler (2013) propose a more generalizable parametric and nonparametric stochastic frontier estimator for count data that is based on a conditional mixed Poisson distribution. Drivas, Economidou, and Tsionas (2014) introduce a Poisson stochastic frontier model that is augmented with a finite mixture structure to allow for heterogeneity of technology classes and that is also able to account for endogeneity of regressors.

Meanwhile, deterministic frontier modelling, which is the focus of the present chapter, originates in the work of Farrell (1957), and comprises two broad subcategories, namely free disposal hull (FDH) and data envelopment analysis (DEA) methods. Deprins, Simar, and Tulkens' (1984) FDH estimator is anchored in the assumption that if $(x, y) \in \Psi$, then for any $x' \geq x$ and $y' \leq y$, $(x', y') \in \Psi$, while Farrell's (1957) and Charnes, Cooper, and Rhodes' (1978) DEA estimator makes an additional assumption of convexity of the production set. Asymptotic results for the FDH and DEA estimators have been derived by Gijbels, Mammen, Park, and Simar (1999) and Park, Simar, and Weiner

(2000), respectively. A comprehensive survey of existing deterministic frontier analysis methods can be found in Simar and Wilson (2013). Unfortunately, as Cazals, Florens, and Simar (2002) have pointed out, deterministic frontier estimation techniques are hindered by their inordinate sensitivity to outliers and so, as an alternative, the authors propose an expected maximal output function that trims the most extreme-valued observations out of the picture. Aragon, Daouia, and Thomas-Agnan (2005) and Martins-Filho and Yao (2008) offer an alternative robust estimation procedure that is based on the conditional quantile function of the output variable. However, an important shortcoming of the aforementioned approaches is their reliance on trimming parameters whose values must typically be selected in an ad hoc fashion. In contrast, the frontier estimator that is proposed in this chapter entails a fairly straightforward and effective trimming parameter selection routine.

The remainder of the chapter is structured as follows: Section 3.2 begins with an overview of robust (deterministic) frontier estimation and then introduces a new approach that can be applied in settings that involve a count-valued output variable and a handful of extreme-valued observations in the data. Section 3.3 shows how one can make use of k-means and higherarchical clustering to determine suitable trimming parameter values for the proposed frontier estimator. Section 3.4 gives an outline of the nonsmooth robust frontier estimator's smooth kernel-based counterpart. Section 3.5 comprises a Monte Carlo experiment that assesses the smooth and nonsmooth estimators' finite-sample performance, while Section 3.6 offers an empirical example involving firm-level patent count data from the U.S. manufacturing sector. Section 3.7 concludes.

3.2 Robust nonparametric frontier estimation

We lay the foundation for the empirical framework that is developed in this chapter by recalling the definition of the conditional distribution of the output variable y given a quantity of inputs x:

$$F\left(y/x\right) = \frac{F\left(y,x\right)}{F_X\left(x\right)}, \tag{3.1}$$

where $F\left(y,x\right) = \mathbb{P}(Y \leq y, X \leq x)$ denotes the joint distribution of the random vector (Y,X), and $F_X(x) = \mathbb{P}(X \leq x)$ is the associated marginal distribution of X.[2] Note that if $x = (x_1, ..., x_d)$ is a $d \times 1$ vector, the inequality $X \leq x$ is shorthand for $(X_1 \leq x_1, ..., X_d \leq x_d)$. Assuming that the conditional distribution function in (3.1) is monotone nonincreasing on the set $\{x : F_X(x) > 0\}$, the production frontier is defined as follows:

$$g(x) = \inf\left\{y : F\left(y/x\right) = 1\right\}. \tag{3.2}$$

Thus, when the function $g(\cdot)$ is evaluated at the vector x, it returns the smallest output value y such that no productive unit with inputs less than or equal to x is able to produce more than y. That is, if we define the order-α conditional quantile for the distribution function that appears in (3.1) as $q_\alpha(x) = F^{-1}(\alpha/x) = \inf\left\{y : F(y/x) \geq \alpha\right\}$ for $\alpha \in [0,1]$, then the frontier in (3.2) can be equivalently expressed as the order-one conditional quantile $q_1(x)$.

Now that the basic framework for deterministic frontier analysis has been

[2]This chapter follows the notational convention whereby a distinction is drawn between the conditional probabilities $F(y/x) = \mathbb{P}(Y \leq y | X \leq x)$ and $F(y|x) = \mathbb{P}(Y \leq y | X = x)$.

briefly described, we move on to a discussion of two important empirical considerations that may arise in applied settings. In particular, in the discussion that appears in Sections 3.2.1 and 3.2.2, we address the challenge of estimating a frontier function that is i) robust to outliers and ii) suitable for models that involve count (as opposed to continuous) data for the output y. Even though there already exist a number of frontier estimators that are robust to the presence of extreme-valued observations - most notably those proposed by Cazals et al. (2002), Aragon et al. (2005), and Martins-Filho and Yao (2008) - it will be shown that they each have their own shortcomings, and hence there is still room for improvement. Meanwhile, to the best of the author's knowledge, this chapter is the first in the deterministic frontier literature[3] to specifically address the issue of robust estimation in a count data setting. It turns out that the manner in which some of the existing approaches deal with the problem of outliers limits their applicability to models in which the output variable is expressed as a count, and therefore, the estimation procedure that is proposed in Section 3.2.2 makes an important contribution to the literature in this regard.

3.2.1 Existing robust estimators

The production frontier in (3.2) is straightforwardly estimated as the order-one empirical quantile of output y conditional on inputs x. That is, for a given sample $\{(x_i, y_i)\}_{i=1}^n$, we compute $\hat{F}(y/x) = \frac{\hat{F}(y,x)}{\hat{F}_X(x)}$, where $\hat{F}(y,x) = \frac{1}{n}\sum_{i=1}^n 1(y_i \leq y, x_i \leq x)$ and $\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^n 1(x_i \leq x)$, and define the

[3]As mentioned in the introduction, Fe (2013), Hofler and Scrogin (2008), Fe and Hofler (2013), and Drivas et al. (2014) have proposed estimators for *stochastic* frontier models involving count data, but their framework is entirely different from the one that is developed here.

frontier estimator as $\hat{g}(x) = \hat{q}_1(x) = \inf\left\{y : \hat{F}(y/x) = 1\right\}$. As mentioned earlier, one of the challenges that is encountered in this setting is that outliers (i.e, observations with unusually large values of y_i) tend to exert too much influence on the estimates of $q_1(x)$. To remedy this issue, Cazals et al. (2002) propose an expected maximum output function of order m:

$$g_m(x) = E(\max(Y_1, ..., Y_m)|X \leq x). \tag{3.3}$$

The order-m frontier estimator is formulated as follows:

$$\begin{aligned} \hat{g}_m(x) &= \int_0^\infty (1 - [\hat{F}(z/x)]^m)dz \\ &= \hat{g}(x) - \int_0^{\hat{g}(x)} [\hat{F}(z/x)]^m dz, \end{aligned} \tag{3.4}$$

where $\hat{g}(x)$ is the order-one conditional output quantile associated with the conditional distribution $\hat{F}(y/x)$ and the trimming parameter m is user-selected.[4] The advantage of this approach is that it is fairly robust to outliers while nonetheless possessing the asymptotic property that $\hat{g}_m(x) \to \hat{g}(x)$ as $m \to \infty$; that is, the estimator is asymptotically equivalent to the empirical order-one conditional quantile $\hat{q}_1(x)$ even though it doesn't envelop all of the data under finite samples. In practice, (3.4) is computed via either numerical integration or a Monte Carlo algorithm that is delineated by Daraio and Simar (2005).

An alternative estimation procedure proposed by Aragon et al. (2005) ensures robustness to extreme output values by defining the production threshold

[4]Cazals et al. (2002) provide a formal proof that if $Y_{max} = \max(Y_1, ..., Y_m)$, then $\mathbb{P}(Y_{max} \leq z|X \leq x) = 1 - [F(z/x)]^m$. Hence, $E(\max(Y_1, ..., Y_m)|X \leq x) = \int_0^\infty (1 - [F(z/x)]^m)dz$. The second equality in (3.4) comes from the fact that $F(z/x) = 1$ for all $z \geq g(x)$. Thus, $\int_0^\infty (1-[F(z/x)]^m)dz = \int_0^{g(x)} (1-[F(z/x)]^m)dz = g(x) - \int_0^{g(x)} [F(z/x)]^m dz$.

in terms of the order-α conditional quantile:

$$q_\alpha(x) = F^{-1}(\alpha/x) = \inf\{y : F(y/x) \geq \alpha\}, \tag{3.5}$$

where $\alpha \in [0,1]$ is the share of productive units with inputs less than or equal to x and output not exceeding y. Thus, one can estimate a frontier function that is less sensitive to outliers than the order-one conditional quantile by setting α to a value that is slightly less than one. In practice, one often chooses a quantile that lies somewhere in the interval $0.90 \leq \alpha \leq 0.99$. Aragon et al.'s (2005) estimator is implemented as follows: for a given input quantity x, one draws a subsample $\{(x_i, y_i)\}_{i=1}^{n_x}$ comprising the n_x observations that satisfy $x_i \leq x$, which gives rise to the distribution function $\hat{F}(y/x) = \frac{1}{n_x}\sum_{i|x_i \leq x}^{n_x} 1(y_i \leq y)$. The production frontier of order α is then estimated as $\hat{q}_\alpha(x) = \min\left\{y_i : \hat{F}(y_i/x) \geq \alpha\right\}$, implying that as $\alpha \to 1$, $\hat{q}_\alpha(x)$ converges to the FDH estimator. Martins-Filho and Yao's (2008) approach is also based on equation (3.5); however, it makes use of a smooth kernel-based estimator of the conditional distribution $F(y/x)$ that has more favourable properties than its nonsmooth empirical counterpart in finite samples. Within this framework, the joint distribution function that appears in (3.1) is estimated as $\hat{F}(x,y) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^y \kappa\left(\frac{y_i - \gamma}{h}\right) d\gamma 1(x_i \leq x)$, where $\kappa(\cdot)$ is a univariate kernel for continuous data (i.e. Gaussian, Epanechnikov) and h is a bandwidth parameter. Thus, the estimator of the order-α production frontier $\hat{q}_\alpha(x) = \hat{F}^{-1}(\alpha/x)$ will always be a smooth function of α, whereas the empirical estimator of Aragon et al. (2005) will not.

At this juncture, it is worth drawing attention to two key limitations of

the frontier estimators that have just been described. First, the trimming procedures that they rely on to curtail the influence of extreme values in the data often have the undesirable side-effect of cutting non-outlier observations out of the picture as well. That is, when either the order-m or the α-quantile methods are being employed, one is faced with a tradeoff whereby sensitivity to outliers can only be reduced at the expense of having a production frontier that does not envelop all of the data. Figures 3.1 and 3.2 illustrate this tradeoff quite clearly. The four plots that appear in the former display a collection of order-m frontier estimates (m=1,2,5,10) obtained for a simulated dataset with n=500 observations, 1 percent of which are outliers.[5] Given that $E(\max(Y_1, ..., Y_m)|X \leq x)$ is decreasing in m, the plots with the headings $m = 1$ and $m = 2$ are characterized by a considerable amount of trimming, whereas in the plots that correspond to $m = 5$ and $m = 10$, a far greater proportion of the data is enveloped by the frontier estimates. Thus, as the value of m increases (decreases), the order-m frontier estimator trims out fewer (more) of the observations and consequently becomes more (less) sensitive to outliers. Meanwhile, the four plots that appear in figure 3.2 depict α-quantile frontier estimates for the same simulated dataset, where α assumes the values 0.99, 0.98, 0.97, and 0.96. In this scenario, a decrease in the quantile parameter α has the same effect as a decrease in the value of m in the previous example, namely more extensive trimming, which reduces the sensitivity of the frontier estimates to extreme-valued observations but also at times leads to inferior envelopment of the data. This is seen in the bottom-right plot that corresponds to $\alpha = 0.96$. On the other hand, in the top-left plot with the heading

[5]This simulated dataset also underlies the second Monte Carlo experiment that is carried out in section 3.5.1.

$\alpha = 0.99$, the frontier estimates are being driven by a small number of outliers and hence they are not very different from what would be obtained under the basic free disposal hull (FDH) framework. One must choose either $\alpha = 0.98$ or $\alpha = 0.97$ in order to obtain an acceptable fit of the model. Ideally, we would like to specify a deterministic frontier estimator that is just as, if not more, robust to outliers as the order-m and α-quantile approaches, but that doesn't require aggressive trimming of the data. This challenge will be addressed in Section 3.2.2.

An additional limitation of Cazals et al.'s (2002) and Martins-Filho and Yao's (2008) methods is that neither one of them is suitable for models in which the output variables is expressed as a discrete count.[6] For instance, closer examination of (3.4) reveals that there is no guarantee that $\hat{g}_m(x) \in \mathbb{Z}_+$, even if $y_i \in \mathbb{Z}_+$ for $i = 1, ..., n$. That is, if the n-observation sample of output quantities has been drawn from a subset of the positive integers, the order-m frontier $\hat{g}_m(x)$ that is obtained by subtracting the real-valued $\int_0^{\hat{g}(x)} [\hat{F}(z/x)]^m dz$ from the positive integer-valued $\hat{g}(x)$ may often belong to $\mathbb{R}_+ \setminus \mathbb{Z}_+$. Thus, in count data settings, using a trimming procedure like the one in (3.4) curtails the influence of outliers but at the expense of generating estimates of the output frontier that cannot, by definition, be elements of the production set. In a similar vein, the nonparametric α-quantile method of Martins-Filho and Yao (2008) is incompatible with a count-valued output variable since it uses a continuous Gaussian or Epanechnikov kernel to estimate the conditional distribution $F(y/x)$. Given that $\hat{F}(y/x)$ is a continuous function of y, when the order-α frontier is estimated by means of the inversion formula

[6]Note that this is not a shortcoming of the empirical α-quantile estimator of Aragon et al. (2005).

$\hat{q}_\alpha(x) = \hat{F}^{-1}(\alpha/x)$, it will frequently be the case that $\hat{q}_\alpha(x) \in \mathbb{R}_+ \setminus \mathbb{Z}_+$. Thus, there is an apparent need for a nonparametric estimator that is i) robust to the presence of outliers, ii) more restrained in its use of trimming, and iii) able to generate exclusively count-valued estimates of the production frontier. In the section that follows, we introduce a frontier estimator that satisfies all three of these criteria.

3.2.2 A robust frontier estimator for count data

The methodological framework that we develop here combines features of both Cazals et al.'s (2002) order-m and Aragon et al.'s (2005) α-quantile definitions of a production frontier, and hence the method that we are proposing shall be referred to as the α-quantile order-m estimator. In short, our analysis focuses on the conditional distribution of $\max(Y_1, ..., Y_m)$ given $X \leq x$, rather than on its conditional mean $E(\max(Y_1, ..., Y_m)|X \leq x)$ which, it will be shown, makes it possible to significantly reduce the influence of outliers without trimming very many (if any) non-outlier observations out of the data. To begin, we delineate the procedure to obtain the α-quantile order-m frontier estimator for values of α and m that are given for now but that will need to be determined at a later stage. The procedure can be described in terms of the following resampling scheme:

1. For $b = 1, ..., B$, draw a subsample of size $m < n$ without replacement $\left\{x_i^{(b)}, y_i^{(b)}\right\}_{i=1}^{m}$
2. Compute the order-one conditional quantile as $\hat{q}_{1,m}^{(b)}(x) = \max\left\{y_i^{(b)} : x_i^{(b)} \leq x\right\}$
3. Define the frontier $\hat{g}_{\alpha,m}(x)$ as the α-quantile of the $\hat{q}_{1,m}^{(b)}(x)$

Thus, for given values of α and m, when the resampling scheme comprises B iterations, the frontier estimator is given by:

$$\hat{g}_{\alpha,m}(x) = \inf \left\{ \hat{q}_{1,m}^{(b)}(x) : \frac{1}{B} \sum_{b'=1}^{B} 1[\hat{q}_{1,m}^{(b')}(x) \leq \hat{q}_{1,m}^{(b)}(x)] \geq \alpha \right\} \tag{3.6}$$

The sequence of steps that is set out above is very similar to the Monte Carlo algorithm that is provided in the appendix of Cazals et al. (2002). Its key distinguishing feature is found in step 3 and equation (3.6), where the α-quantile rather than the mean of the $\hat{q}_{1,m}^{(b)}(x)$ is computed. Note that there are two separate justifications for following this approach. First, it is very likely that some, and perhaps many, of the B iterations of the resampling scheme will assign outlier values of y_i to $\hat{q}_{1,m}^{(b)}(x)$, and this will bias one's estimate of the conditional expectation $E(\max(Y_1, ..., Y_m)|X \leq x)$. In contrast, if a suitable value of α is used in step 3 and equation (3.6) above, then the estimator $\hat{g}_{\alpha,m}(x)$ will not suffer from this bias. Second, replacing the conditional mean of $\max(Y_1, ..., Y_m)$ given $X \leq x$ with its α-quantile ensures that the frontier function always returns a count value. As mentioned in Section 3.2.1, in the event that the output variable $y \in \{0, 1, ..., p-1\}$ can only be sensibly expressed as one of p possible non-negative integers, the standard order-m trimming procedure will often give rise to frontier estimates that lie in $\mathbb{R}_+ \setminus \mathbb{Z}_+$. This will not occur if the frontier function is defined in terms of $\hat{g}_{\alpha,m}(x)$.

Of course, without detailed knowledge of the data generating process and by extension, the nature of the outlier problem, there is nothing that immediately suggests what values of the trimming parameters α and m might be appropriate for robust estimation of the frontier. As $\alpha \to 1$ one $m \to n$,

$\hat{g}_{\alpha,m}(x)$ converges to the FDH (i.e. the order-one quantile) estimator, and consequently becomes more sensitive to extreme-valued observations in the data. Figure 3.3 provides an intuitive illustration of this tendency. Using the same simulated dataset with n = 500 (1 percent of which are outliers) that was referenced in the previous subsection, the figure comprises four plots that each correspond to a particular combination of α and m. The two plots on the right-hand side indicate that assigning too great a value to either parameter when outliers are present in the data can lead to very poor estimates of the frontier function; in fact, these estimates are more or less indistinguishable from what one would obtain under the FDH approach. In contrast, the two plots on the left-hand side of figure 3.3 demonstrate that the right amount of trimming (in this instance, either $\alpha = 0.5$ and $m = 50$ or $\alpha = 0.05$ and $m = 200$) yields improvements over the standard order-m and α-quantile frontiers that are depicted in figures 3.1 and 3.2. For the sake of comparison, figure 3.4 comprises frontier plots that are based on simulated data without any outliers; in this setting, most values of the trimming parameters, provided they are not too close to their respective lower bounds, produce sensible estimates of the frontier function, although $\alpha = 1$ and $m = n$ are obviously optimal. Therefore, under the typical real-world scenario of an unknown DGP, the challenge is to select values of α and m that will trim outliers out of the data while still ensuring that $\hat{g}_{\alpha,m}(x)$ envolops as many non-outlier observations as possible. We now move on to a discussion of how this objective can be achieved.

3.3 Cluster-based selection of trimming parameters

The α-quantile order-m framework requires a preliminary assessment of the sensitivity of the frontier estimates to the chosen values of the trimming parameters. To make this determination, we follow a clustering approach that is somewhat computationally-demanding, but that can also lead to substantial improvements over existing trimming methods. Given that the proposed estimator $\hat{g}_{\alpha,m}(x)$ depends on α and m, if we have a sample of size n, we can fit a frontier function as in (3.6) over a grid of combinations of $\alpha = \alpha_{min}, ..., 1$ and $m = m_{min}, ..., n$, where α_{min} and m_{min} denote the parameters' pre-established lower bounds. We consider the same simulated example as in the previous section where the dataset comprises $n = 500$ observations, 1 percent of which are outliers, and where the frontier is specified as a step function. We apply our estimation procedure for every pairwise combination of $\alpha = 0.05, 0.10, ..., 0.95, 1$ and $m = 30, 40, ..., 490, 500$, which implies 960 different fits of the frontier model. It should take a few minutes at most for a computer with a reasonably fast processor to perform this operation. Given that the input $x \in \{1, 2, ..., 10\}$ is modelled as a count with 10 unique values, each of the 960 points in the grid will have a corresponding 10×1 fitted vector $\hat{\mathbf{g}}_{\alpha,m} = [\hat{g}_{\alpha,m}(1), ..., \hat{g}_{\alpha,m}(10)]'$.[7] If outliers have a lot of leverage over the frontier estimates, then the vector of the $\hat{g}_{\alpha,m}(x)$ should vary substantially across different pairwise combinations of the trimming parameters. In particular, one should be able to divide the frontier estimates into two disjoint clusters - one with relatively large values

[7] In the event that $x = (x_1, ..., x_d)$ is multivariate with a mixture of continuous and discrete count data, then if one defines n_e evaluation points for each of $x_1, ..., x_d$, the fitted vector $\hat{\mathbf{g}}_{\alpha,m}$ will be $n_e^d \times 1$ rather than 10×1. In practice, one could set $n_e = 5$ and use the 0.10^{th}, 0.30^{th}, 0.50^{th}, 0.70^{th}, and 0.90^{th} quantiles of each of $x_1, ..., x_d$.

of α and m that displays sensitivity to extreme observations and another with smaller α and m that is more robust to outliers. Let these two clusters of frontier estimates be denoted by C_1 and C_2. As explained in Hastie, James, Tibshirani, and Witten (2013), one needs to solve the minimization problem below:

$$\min_{C_1,C_2} \left\{ \sum_{k=1}^{2} \frac{1}{|C_k|} \sum_{\substack{(\alpha,m)\in C_k \\ (\tilde{\alpha},\tilde{m})\in C_k}} \hat{\mathbf{g}}'_{\alpha,m} \hat{\mathbf{g}}_{\tilde{\alpha},\tilde{m}} \right\}, \tag{3.7}$$

where $|C_k|$ denotes the number of different $\hat{\mathbf{g}}_{\alpha,m}$ in cluster k. Hence the frontier estimates that are most similar to one another will end up in the same cluster. For details on the algorithm that is used to solve (3.7), the reader is encouraged to consult Hastie et al. (2013), or the documentation for the `kmeans` function in the R statistical computing environment.

Figure 3.5 plots the outcome of this k-means clustering exercise for the simulated dataset. The separation pattern of the frontier estimates based on the values of α and m used to compute them is indicative of a potential outlier problem. For the sake of juxtaposition, we also apply the k-means clustering method to data that does not include any outliers. The outcome is plotted in figure 3.6. In this instance, we find that the lower-left cluster of frontier estimates shrinks to a fraction of its original size; only the very smallest values of α and m give rise to $\hat{\mathbf{g}}_{\alpha,m}$ that can be considered meaningfully different from the rest.[8] This limited degree of separation suggests that the frontier is not particularly sensitive to the choice of trimming parameters, and consequently, the FDH estimator might be appropriate after all.

[8] In fact, given that very small values of α and m will always be grouped together because they tend to collectively understate the magnitude of the frontier, (3.7) can be specified as a 3-cluster problem, where the cluster containing $\hat{\mathbf{g}}_{\alpha_{min},m_{min}}$ is ultimately discarded.

k-means clustering facilitates outlier detection to the extent that it identifies fundamental differences between $\hat{g}_{\alpha,m}(x)$ and the FDH estimator for various choices of α and m. However, it does not provide a clear suggestion of which values of α and m ought to be used in (3.6), since not all combinations of the trimming parameters that are assigned to the lower-left cluster in figure 3.5 yield frontier estimates that are 100 percent robust to outliers. Interestingly, hierarchical clustering can serve as a very useful guide when choosing among possible combinations of the trimming parameters. If, after performing the k-means procedure, one suspects that the frontier estimates are indeed being influenced by a handful of outliers, then suitable values of α and m are likely to be found somewhere in the lower-left portion of figure 3.5. We take into consideration all of the $\hat{\mathbf{g}}_{\alpha,m}$ that satisfy $\alpha \leq 0.5$ and $m \leq 200$. If we perform hierarchical clustering on this subset of the $\hat{\mathbf{g}}_{\alpha,m}$, we can plot the results in a dendrogram such as the one that appears in figure 3.7. The interested reader is encouraged to consult Hastie et al. (2013) for a thorough explanation of hierarchical clustering and interpretation of dendrogram plots. In short, each of the observations (i.e. the subset of the $\hat{\mathbf{g}}_{\alpha,m}$) is represented by its own branch at the base of the dendrogram, and the hierarchical procedure progressively clusters the vectors of frontier estimates in the order of their similarity to one another. The distance that separates two observations or clusters is given by the point on the vertical axis at which they are fused together. Hence, we should expect the robust frontier estimates to form a cluster at or very near the base of the dendrogram, that is, somewhere in the neighbourhood of zero on the vertical axis. Once this cluster has been identified, one can choose any of its associated pairs of α and m. For instance, in figure 3.7, there is a very

large cluster of frontier estimates at the centre-left of the base of the dendrogram. If we select a combination of α and m that corresponds to any one of the fused branches, we will find that $\hat{\mathbf{g}}_{\alpha,m}$ ends up being a perfect estimate of the frontier.[9] Further experimentation in Section 3.5 with this hierarchical clustering procedure for trimming parameter selection reveals that the proposed α-quantile order-m frontier estimator can substantially improve upon existing robust methods in finite sample settings. Before elaborating on this point, however, we first describe an extension of our robust nonparametric estimator that relies on kernel smoothing instead of an empirical conditional distribution function.

3.4 A smooth frontier estimator for count data

In the preceding discussion, we proposed a robust nonparametric frontier estimator that was based on a nonsmooth (i.e. empirical) conditional distribution function $\hat{F}(y/x) = \frac{\hat{F}(y,x)}{\hat{F}_X(x)}$. However, as Martins-Filho and Yao (2008) and others have pointed out, it might be advantageous to use a smooth kernel-based estimator instead. Hence we now consider a setting in which output $y \in \{0, 1, ..., p-1\}$ is expressed as one of p possible count values, and where the $(q+r)$-dimensional input vector $x = (x^c, x^d)$ comprises q continuous and r discrete count variables, respectively denoted by $x^c = (x_1^c, ..., x_q^c)$ and $x^d = (x_1^d, ..., x_r^d)$. The objective is to estimate a frontier function that is based

[9] This procedure can be automated via identification of the largest cluster for a particular cut-off point on the vertical axis of the dendrogram; however, the optimal cut-off value will be context-specific insofar as it depends on the range and variability of the frontier estimates for different α and m. In the present example, using the base of the dendrogram as a cut-off point is the best option, but this will not always be the case. Hence the clustering method is primarily an exploratory tool whose interpretation must ultimately be left to the researcher.

on the conditional distribution $F(y/x) = \frac{F(y,x)}{F_X(x)}$, which is now defined as:

$$F(y/x) = \frac{F(y,x)}{F_X(x)} = \frac{\mathbb{P}(Y \leq y, X^c \leq x^c, X^d \leq x^d)}{\mathbb{P}(X^c \leq x^c, X^d \leq x^d)}. \tag{3.8}$$

Inversion of (3.8) gives rise to a FDH-style frontier function that is analogous to the one in (3.2), where $g(x) = q_1(x) = \inf\{y : F(y/x) = 1\}$ now reflects the fact that y is a count and that the vector x comprises a mix of continuous and discrete count data.

Estimation of (3.8) can proceed along the lines of the kernel-based framework that is laid out in Li and Racine (2003, 2008). This particular nonparametric estimator for distributions that involve mixed continuous and discrete data is ideally suited to the present setting, since we would like to model a count output frontier that is a function of a mixed-data input vector. We begin by defining a kernel weighting function for the count output y that can also be applied to each of the r discrete count inputs in the vector $x^d = (x_1^d, ..., x_r^d)$:

$$l_\lambda(y_i, y) = \begin{cases} 1 & \text{if } y_i = y \\ \lambda^{|y_i - y|} & \text{otherwise ,} \end{cases} \tag{3.9}$$

where $\lambda \in [0, 1]$ is a smoothing parameter whose value is yet to be determined. In regard to the continuous inputs $x^c = (x_1^c, ..., x_q^c)$, we follow a similar approach to that of Martins-Filho and Yao (2008) and define $w_{h_j}(x_{ij}^c, x_j^c) = \frac{1}{h_j}\kappa\left(\frac{x_{ij}^c - x_j^c}{h_j}\right)$ for $j = 1, ..., q$, where h_j is a bandwidth parameter and the bounded, symmetric kernel $\kappa(\cdot)$ satisfies $\int \kappa(z)dz = 1$, $\int z^2\kappa(z)dz > 0$, and $\int \kappa^2(z)dz > 0$. In the analysis that follows, a second-order Gaussian kernel is used for $\kappa(\cdot)$. Finally, the weighting function for the discrete count input x_k^d,

where $k = 1, ..., r$, is defined exactly as in (3.9) and is henceforth denoted by $l_{\gamma_k}(x_{ik}^d, x_k^d)$ for a given smoothing parameter $\gamma_k \in [0, 1]$.

Li and Racine (2003) introduce the notion of a "general product kernel" so that weights can be assigned to observations in the sample based on the $(q + r)$-dimensional vector $x = (x^c, x^d)$. Let $W_h(x_i^c, x^c) = \prod_{j=1}^{q} w_{h_j}(x_{ij}^c, x_j^c)$ and $L_\gamma(x_i^d, x^d) = \prod_{k=1}^{r} l_{\gamma_k}(x_{ik}^d, x_k^d)$ denote product kernels for the continuous and discrete count variables, respectively. The nonparametric estimator of the joint density function $f(y, x)$ is given by:

$$\hat{f}(y, x) = \frac{1}{n} \sum_{i=1}^{n} l_\lambda(y_i, y) \times W_h(x_i^c, x^c) \times L_\gamma(x_i^d, x^d). \tag{3.10}$$

In a similar vein, the estimator of the marginal density function $f_X(x)$ is given by:

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^{n} W_h(x_i^c, x^c) \times L_\lambda(x_i^d, x^d). \tag{3.11}$$

Estimators of the joint and marginal distributions are then obtained by plugging the density functions in (3.10) and (3.11), respectively, into the expressions below:

$$\hat{F}(y, x) = \sum_{y' \leq y} \sum_{x_1^{o\prime} \leq x_1^o} \cdots \sum_{x_r^{o\prime} \leq x_r^o} \int_0^{x_1^c} \cdots \int_0^{x_q^c} \hat{f}\left(x_1^{c\prime}, ..., x_q^{c\prime}, x_1^{d\prime}, ..., x_r^{d\prime}, y'\right) dx_1^{c\prime} \cdots dx_q^{c\prime} \tag{3.12}$$

$$\hat{F}_X(x) = \sum_{x_1^{d\prime} \leq x_1^d} \cdots \sum_{x_r^{d\prime} \leq x_r^d} \int_0^{x_1^c} \cdots \int_0^{x_q^c} \hat{f}_X\left(x_1^{c\prime}, ..., x_q^{c\prime}, x_1^{d\prime}, ..., x_r^{d\prime}\right) dx_1^{c\prime} \cdots dx_q^{c\prime}. \tag{3.13}$$

Thus, the joint and marginal CDFs are obtained by either summing or integrating the joint and marginal density functions over all of the input and output variables. Fortunately, (3.12) and (3.13) have straightforward analytical solu-

tions because the integral is well-defined for the second-order Gaussian kernel appearing in the local weighting functions for the q continuous inputs. In particular, the integral of $W_h(x_i^c, x^c)$ in (3.12) and (3.13) is given by $G_h(x_i^c, x^c) = \int_0^{x_1^c} \cdots \int_0^{x_q^c} \left(\prod_{j=1}^q w_{h_j}(x_{ij}^c, x_j^{c\prime}) \right) dx_1^{c\prime} \cdots dx_q^{c\prime} = \prod_{j=1}^q \frac{1}{h_j} \int_0^{x_j^c} \kappa \left(\frac{x_{ij}^c - x_j^{c\prime}}{h_j} \right) dx_j^{c\prime}$, while the $(r+1)$ summation expressions that involve the discrete count variables' weighting functions $l_\lambda(y_i, y)$ and $L_\gamma(x_i^d, x^d) = \prod_{k=1}^r l_{\gamma_k}(x_{ik}^d, x_k^d)$ are rather easy to compute.

The final task that remains before we can proceed with estimation of the distribution functions in (3.12), (3.13), and (3.1) is to select appropriate values of the $q+r+1$ smoothing parameters λ, $h = (h_j)_{j=1}^q$, and $\gamma = (\gamma_k)_{k=1}^r$, which are henceforth denoted by the shorthand $\theta = (\lambda, h, \gamma)$. Li, Li, and Racine (2014) build on previous work by Ju, Li, and Liang (2009) and propose a data-driven bandwidth selection method for CDFs that involve a combination of continuous and discrete variables. Under this framework, we first generate a set of evaluation points $\{x_j, y_j\}_{j=1}^{n_e}$ by drawing a subsample of size $n_e < n$ (without replacement) from the n-observation dataset $\{x_i, y_i\}_{i=1}^n$. We then use a nonlinear optimization routine to identify the vector of smoothing parameters θ that minimizes the cross-validation function below:

$$CV(\theta) = \frac{1}{nn_e} \sum_{i=1}^{n} \sum_{j=1}^{n_e} [1(y_i \leq y_j, x_i^c \leq x_j^c, x_i^d \leq x_j^d) - \hat{F}_{-i}(y_j, x_j^c, x_j^d)]^2, \quad (3.14)$$

where $\hat{F}_{-i}(y_j, x_j^c, x_j^d)$ is the "leave-one-out" estimator of the joint distribution function. It is obtained by deleting observation i from the sample, using the $n-1$ observations that remain to compute (3.10) and (3.12), and then evaluating the function at (y_j, x_j^c, x_j^d) for $i = 1, ..., n$ and $j = 1, ..., n_e$. Note that

(3.14) applies to the joint distribution $F(\cdot)$, and hence one has the option of either using the h and γ that minimize $CV(\theta)$ above to estimate the marginal CDF $F_X(\cdot)$, or repeating the bandwidth selection procedure for a new cross-validation function that excludes the output variable y. In this chapter, we opt for the latter.

Now that we have established how the nonparametric conditional distribution estimator $\hat{F}(y/x) = \frac{\hat{F}(y,x)}{\hat{F}_X(x)}$ can be defined along the lines of Li and Racine's (2003; 2008) and Li et al.'s (2014) kernel-based framework, we can model the α-quantile order-m frontier function for a count ouput variable in more or less the same fashion as in Section 3.2.2. The procedure is delineated as follows:

1. For $b = 1, ..., B$, draw a subsample of size $m < n$ without replacement $\left\{x_i^{(b)}, y_i^{(b)}\right\}_{i=1}^{m}$
2. Estimate the smoothed conditional distribution $\hat{F}^{(b)}(y/x) = \frac{\hat{F}(y,x)}{\hat{F}_X^{(b)}(x)}$
3. Compute the order-one conditional quantile $\hat{q}_1^{(b)}(x) = \inf\left\{y : \hat{F}^{(b)}(y/x) = 1\right\}$
4. Define the frontier function as the order-α quantile of the $\hat{q}_1^{(b)}(x)$

Selection of suitable values of α and m can once again be undertaken using the k-means and hierarchical clustering methods that were outlined in Section 3.2.2. By now, it should be obvious that the smooth frontier estimator for count data is asymptotically-equivalent to the FDH framework in (3.2) insofar as $\hat{g}_{\alpha,m}(x)$ converges to the order-one conditional quantile $\hat{q}_1(x)$ as $\alpha \to 1$ and $m \to n$. In this sense, it mimics the large-sample behaviour of Cazals et al.'s (2002) order-m estimator and Aragon et al.'s (2005) α-quantile method. Note that the smooth conditional distribution function $\hat{F}(y/x)$ that is obtained using Li and Racine's (2003; 2008) general product kernel can also be applied to

the formulation of a count data analogue of Martins-Filho and Yao's (2008) α-quantile frontier estimator. However, we leave this extension of the present framework as a topic for future research. In the section that follows, we consider a number of simulated examples that shed light on both the smooth and the nonsmooth α-quantile order-m estimators' behaviour in finite-sample settings. It turns out that when the frontier model is characterized by i) a count output variable and ii) outliers in the data, both of the estimators that we have proposed consistently outperform existing nonparametric approaches.

3.5 Monte Carlo simulation

We now perform a Monte Carlo experiment in order to evaluate the proposed α-quantile order-m estimator's performance under a number of different specifications of the frontier model. In a simulated setting where the production set's underlying data generating process is known, we would like to compare the results of the estimation procedure in Sections 3.2.2 and 3.4 with those of existing nonparametric approaches. In the first part of what follows, we use the estimators of Cazals et al. (2002) and Aragon et al. (2005), respectively, as benchmarks for comparison. The key attribute that ties these two approaches together is their robustness to outliers by means of trimming methods that lead to a certain proportion of the data points not getting enveloped by the frontier. Hence they form a relevant backdrop against which the finite-sample behaviour of our nonsmooth α-quantile order-m frontier estimator for count data can be assessed - specifically, we would like to establish whether the newly-proposed trimming procedure offers meaningful improvements over methods that already exist. As an extension, we also incorporate Martins-Filho and

Yao's (2008) kernel-based α-quantile framework into the simulation exercises and contrast its performance under a variety of DGPs with that of our smooth nonparametric frontier estimator for count data. All of the supporting tables and figures that are referenced below can be found in the appendix.

3.5.1 Nonsmooth estimators

In this first experiment, we consider a scenario where the efficient output count is a function of a single discrete count variable. Furthermore, we assume that there are a handful of extreme-valued observations in the data that have the potential to distort the statistical analysis of productive efficiency. We consider three specifications of the non-outlier data generating process for the production set Ψ:

1. $\Psi = \{(X, Y) : Y \leq X\}$
2. $\Psi = \{(X, Y) : Y \leq s(X)\}$
3. $\Psi = \{(X, Y) : Y \leq X^2\}$,

where $s(X)$ in the second instance above is a step function with two more or less equally-spaced jumps. In each instance, the input $X \sim \mathcal{U}\{1, 10\}$ and the output $Y \sim \mathcal{U}\left\{\left\lfloor \frac{g(X)}{2} \right\rfloor, g(X)\right\}$ are drawn from a discrete uniform distribution, where $g(X)$ denotes the frontier that envelops all non-outlier data points for a given value of X,[10] and $\left\lfloor \frac{g(X)}{2} \right\rfloor$ denotes the largest integer that is less than or equal to $\frac{g(X)}{2}$. We consider sample sizes of $n = 500$ and $n = 1000$

[10]Thus, we have $g(X) = X$, $g(X) = s(X)$, and $g(X) = X^2$ under the first, second, and third specifications of the frontier model, respectively. The step function is defined as $s(X) = 3 \cdot 1(1 \leq X \leq 3) + 7 \cdot 1(4 \leq X \leq 7) + 10 \cdot 1(8 \leq X \leq 10)$, where $1(\cdot)$ denotes an indicator function.

where in each case, one percent of the total number of observations are assigned extreme values. One can assume that outliers such as these have arisen due to measurement error or some other imperfection in the data collection process. The outliers in the simulation exercise are generated as follows: for a subset of input values $X \in \{1, 2, ..., 10\}$, we define various output intervals such that there is a significant amount of separation between each interval and the frontier. Hence, the endpoints that characterize each of these intervals will depend on both the value of the input variable X and the specification of the frontier function $g(X)$ (i.e., linear, step, or quadratic). In each iteration of the Monte Carlo experiment, we draw a new set of outlier observations from these pre-determined intervals in order to introduce some variation into the outlier generation process.[11]

We estimate the linear, step, and quadratic frontier models using our proposed nonsmooth nonparametric approach, the order-m method of Cazals et al. (2002) that is depicted in equations (3.3) and (3.4), and the α-quantile framework of Aragon et al. (2005) that is outlined in equation (3.5). Given that we only briefly alluded to the numerical integration procedure that is involved in the order-m estimator's implementation, we provide a quick overview of what it entails before proceeding with the Monte Carlo simulation. For a given $x \in \{1, ..., 10\}$, we let $n_x = \sum_{i=1}^{n} 1(x_i \leq x)$ and draw a subsample of all n_x observations satisfying $x_i \leq x$, which yields the ordered sequence of output

[11] For example, under the quadratic specification of the production set with $X \sim \mathcal{U}\{1, 10\}$, $\Psi = \{(X, Y) : Y \leq X^2\}$, and $n = 500$, we take the subset $x = 3, 4, 5, 6, 7$ and draw the five corresponding output outliers from the intervals $\{45, ..., 55\}$, $\{45, ..., 55\}$, $\{60, ..., 80\}$, $\{50, ..., 60\}$, and $\{70, ..., 90\}$. Note the considerable distance that separates the lower endpoints of these intervals and the respective output frontier values of $g(x) = 9, 16, 25, 36, 49$. Replication code that details the outlier generation process for each specification of the frontier model is available upon request.

values $\{y_j\}_{j=1}^{n_x}$, where $y_1 < y_2 < ... < y_{n_x}$. The order-$m$ frontier $\hat{g}_m(x)$ is given by:

$$\hat{g}_m^{ns}(x) = \sum_{j=2}^{n_x}(1 - [\hat{F}(y_j/x)]^m)(y_j - y_{j-1}), \tag{3.15}$$

where for $j = 1, ..., n_x$, the empirical conditional distribution function is computed as $\hat{F}(y_j/x) = \frac{j}{n_x}$. Meanwhile, the α conditional quantile frontier estimate is given by $y_{\lceil \alpha n_x \rceil}$, where $\lceil \alpha n_x \rceil$ denotes the smallest integer that is greater than or equal to αn_x.

The results of the Monte Carlo experiment are reported in table 3.1. Each entry corresponds to the average root mean squared error (RMSE) for a particular specification of the model (i.e. linear, step, quadratic) and sample size, based on $M = 500$ random draws of the data. The RMSE that is recorded for both the order-m and the α-quantile estimators is the minimum across all values of m and α, respectively.[12] Thus, the first four columns should be interpreted as the "best" possible outcomes for the nonsmooth estimators of Cazals et al. (2002) and Aragon et al. (2005) in this simulated setting; their finite-sample performance is not quite as impressive as one moves away from the optimal m and α. Meanwhile, the proposed α-quantile order-m estimator is implemented using values of α and m that have been selected via the hierarchical clustering method described in Section 3.3, and hence the average RMSE that is reported in the fifth and sixth columns should be close but not necessarily equal to the minimum across all possible combinations of the trimming parameters. Three key results of this exercise are worth highlighting. First,

[12] In particular, the frontier model is estimated using $m = 1, ..., 15$ and $\alpha = 0.94, 0.95, ..., 0.99, 1$ (anything beyond $m = 15$ or $\alpha = 0.94$ results in a very large RMSE due to excessive trimming), and whichever of these values yields the smallest average RMSE is used to compute what appears in the first four columns of table 3.1.

the order-m estimator is consistently outperformed by both the α-quantile and α-quantile order-m methods. We observe that under all three specifications of the frontier model, the average RMSE in columns 1 and 2 is considerably higher than in columns 3 through 6. Second, the α conditional quantile tends to provide very good estimates of the frontier when either $\alpha = 0.98$ or $\alpha = 0.99$. For example, when the frontier is either a linear or a step function of x, the pointwise RMSE associated with $\hat{q}_\alpha(x)$ is remarkably close to zero, although this only holds when one has chosen the optimal value of α. Third and most importantly, the estimator that was proposed in Section 3.2.2 has the lowest RMSE in all six instances that are considered in table 3.1. That is, it actually outperforms $\hat{q}_\alpha(x)$ with an ideally-chosen trimming parameter, and provides near-perfect estimates of the linear and step frontiers.

At this juncture, it is worth pointing out that the results of the Monte Carlo simulation are not intended as a suggestion that our proposed method will *always* give rise to the best fit of the production set; this would indeed be somewhat of an exaggeration. The clustering method that is used for trimming parameter selection should, in general, lead to a better fit than the more rudimentary order-m or α-quantile approaches, but one shouldn't expect this outcome 100 percent of the time. Rather, the preceding discussion has established that for sensibly-chosen values of α and m, the α-quantile order-m frontier estimator can perform just as well, if not better, than the optimal versions of $\hat{g}_m(x)$ and $\hat{q}_\alpha(x)$.

3.5.2 Smooth estimators

In this second exercise, we compare our smooth frontier estimator for count data with the kernel-based α-quantile method of Martins-Filho and Yao (2008). Given that the latter will tend to perform very well when the frontier function exhibits a considerable amount of curvature, we specify the data generating process as follows:

1. $\Psi = \left\{(X, Y) : Y \leq 2^{X-1}\right\}$

2. $\Psi = \{(X, Y) : Y \leq 100 + (X - 5)^3\}$.

The bandwidth selection for the smooth α conditional quantile estimator $\hat{q}_\alpha(x)$ can be carried out using the same cross-validation procedure that appears in (3.14), with the exception that smoothing will in this instance only be applied to the output variable y, which is treated as continuous rather than discrete. Once again, we consider sample sizes $n = 500$ and $n = 1000$, and assume that one percent of the total number of observations consists of outliers that lie well outside of Ψ. The outlier generation process is the same as what was described in the previous subsection, and replication code is available upon request. Considering the amount of variation that is exhibited by the output frontier under the two specifications of the model, table 3.2 suggests that both Martins-Filho and Yao's (2008) approach and the method that we are proposing in this chapter perform quite well. As in the first simulation, the average RMSE that is reported for the smooth α-quantile estimator is the minimum across a number of different values of α. Meanwhile, the fit of the α-quantile order-m estimator is based on trimming parameter values that have been selected via hierarchical clustering - the chosen values of α and m are

reported in the description under table 3.2. Altogether, the approach that is delineated in Section 3.4 gives rise to the lowest RMSE; however, to reiterate what was mentioned in the last paragraph of Section 3.5.1, the objective of this Monte Carlo experiment is not to demonstrate that our procedure will always outperform existing methods. Instead, it merely shows that the proposed smooth estimator $\hat{g}_{\alpha,m}(x)$ is well-behaved in finite sample settings; that is, it yields a fit of the frontier function that is comparable to the smooth version of $\hat{q}_\alpha(x)$ with an optimally-chosen trimming parameter α. On top of this, it has the added benefit that it is fully compatible with a count-valued output variable, unlike the competing smooth estimator that is primarily suited for continuous data types.

3.6 An empirical example using firm-level patent data

In this section of the chapter, we estimate a frontier model in which the dependent variable is a count of patents granted to U.S. manufacturing firms between 1975 and 1979. The dataset that we use originally appeared in Hall, Griliches, and Hausman (1986), and has been made publicly available by Cameron and Trivedi (2013). The five-year balanced panel comprises a total of 735 observations on 147 unique firms. Five different industries are represented in the sample, namely manufacturing of pharmaceuticals, computers, scientific instruments, chemicals, and electric components. To ensure lucidity of exposition, we estimate a very simple model that specifies a firm's patent count as a univariate function of the log of its investment in research and development in any given year. Hence in this context, the frontier estimator ought to be able to shed light on the efficiency with which various firms translate their R&D

expenditures into proprietary technology.

We begin by estimating the α-quantile order-m frontier for a number of different pairwise combinations of the trimming parameters, and we plot the outcome of the k-means clustering outlier detection procedure in figure 3.8. While there is not quite as much separation of the fitted vectors $\hat{\mathbf{g}}_{\alpha,m}$ as there was in the simulated example from Section 3.3, we nevertheless see some evidence of outliers in the data. For the sake of illustration, we estimate the frontier using values of α and m that appear in the lower-left portion of figure 3.8, and contrast this with what is obtained under the free disposal hull framework (i.e., the order-one quantile with m=n). The results of this exercise are plotted in figures 3.9 and 3.10. When $\alpha = 0.3$ and $m = 240$, roughly 7 percent of the observations do not get enveloped by the frontier, unlike when the FDH method is implemented. The efficiency of each firm's R&D endeavours can be computed in terms of its relative distance from the patent count frontier; that is, if x_{it} and y_{it} denote firm i's period-t R&D spending and patent count, respectively, then the α-quantile order-m measure of efficiency is given by $\frac{y_{it}}{\hat{g}_{\alpha,m}(x_{it})}$, whereas the FDH measure is given by $\frac{y_{it}}{\hat{q}_1(x_{it})}$. It should be obvious by now that the inequality $\frac{y_{it}}{\hat{q}_1(x_{it})} \leq \frac{y_{it}}{\hat{g}_{\alpha,m}(x_{it})}$ will always hold because $\hat{g}_{\alpha,m}(x_{it}) \leq \hat{q}_1(x)$ by construction. This relationship can be clearly seen in figure 3.12, where we have plotted the respective distributions of the two categories of efficiency estimates. Note that extreme-valued observations are not as much of a problem in Hall et al.'s (1986) firm-level dataset as they were in some of the simulated examples that were presented earlier in this chapter, and hence the distance that separates the two CDF plots might actually be a lot wider under alternative circumstances. At the end of the day, in applied

settings, the decision of whether or not to resort to data trimming is left to the judgement of the researcher, although whatever the case may be, the α-quantile order-m procedure is a sound benchmark against which the robustness of the FDH frontier can be evaluated.

3.7 Conclusion

This chapter has proposed a robust nonparametric estimation procedure for deterministic frontier models with a count-valued dependent variable. It has argued that a cluster-based approach to trimming parameter selection may be preferable to the ad hoc methods that are commonly employed by existing robust estimators, and it has provided Monte Carlo evidence to support this claim. For suitably-chosen trimming parameter values, both the nonsmooth and the smooth versions of the proposed α-quantile order-m estimator are well-behaved in finite sample settings, and they generally give rise to a better fit of the outlier-free frontier than competing approaches, namely those based on either a conditional quantile or an expected maximal output function. An empirical example has been provided using publicly-available data on firm-level patent counts and R & D spending; in this instance, 7 percent of the observations are labelled as outliers, and are consequently not enveloped by the robust frontier. Refinement of the hierarchical clustering framework so that it allows for more automatic trimming of extreme-valued data points offers many intriguing possibilities for future research.

References

Aigner, D., C. A. K. Lovell, and P. Schmidt (1977): "Formulation and estimation of stochastic frontier production function models," *Journal of Econometrics*, 6, 21–37.

Aragon, Y., A. Daouia, and C. Thomas-Agnan (2005): "Nonparametric frontier estimation: A conditional quantile-based approach," *Econometric Theory*, 21, 358–389.

Cameron, A. C. and P. K. Trivedi (2013): *Regression Analysis of Count Data*, Cambridge,UK: Cambridge University Press.

Cazals, C., J. P. Florens, and L. Simar (2002): "Nonparametric frontier estimation: A robust approach," *Journal of Econometrics*, 106, 1–25.

Charnes, A., W. W. Cooper, and E. Rhodes (1978): "Measuring the inefficiency of decision making units," *European Journal of Operational Research*, 2, 429–444.

Daraio, C. and L. Simar (2005): "Introducing environmental variables in nonparametric frontier models: A probabilistic approach," *Journal of Productivity Analysis*, 24, 93–121.

Debreu, G. (1951): "The coefficient of resource utilization," *Econometrica*, 19, 273–292.

Deprins, D., L. Simar, and H. Tulkens (1984): "Measuring labor inefficiency in post offices," in *The Performance of Public Enterprises: Concepts*

and Measurements, ed. by M. Marchand, P. Pestieau, and H. Tulkens, Amsterdam: North-Holland, 243–267.

Drivas, K., C. Economidou, and E. Tsionas (2014): "A Poisson stochastic frontier model with finite mixture structure," Mpra paper, University Library of Munich, Germany.

Farrell, M. J. (1957): "The measurement of productive efficiency," *Journal of the Royal Statistical Society Series A*, 120, 253–290.

Fe, E. (2013): "Estimating production frontiers and efficiency when output is a discretely distributed economic bad," *Journal of Productivity Analysis*, 39, 285–302.

Fe, E. and R. Hofler (2013): "Count data stochastic frontier models, with an application to the patents-R&D relationship," *Journal of Productivity Analysis*, 39, 271–284.

Gijbels, I., E. Mammen, B. U. Park, and L. Simar (1999): "On estimation of monotone and concave frontier functions," *Journal of the American Statistical Association*, 94, 220–228.

Hall, B. H., Z. Griliches, and J. A. Hausman (1986): "Patents and R and D: Is there a lag?" *International Economic Review*, 27, 265–283.

Hastie, T., G. James, R. Tibshirani, and D. Witten (2013): *An Introduction to Statistical Learning with Applications in R*, New York, NY: Springer.

Hofler, R. A. and D. Scrogin (2008): "A count data frontier model," Tech. rep., University of Central Florida Department of Economics.

Ju, G., R. Li, and Z. Liang (2009): "Nonparametric estimation of multivariate CDF with categorical and continuous data," in *Advances in Econometrics: Nonparametric Econometric Methods*, ed. by Q. Li and J. S. Racine, Elsevier Science, vol. 25, 291–318.

Koopmans, T. C. (1951): "Analysis of production as an efficient combination of activities," in *Activity Analysis of Production and Allocation*, ed. by T. C. Koopmans, New York: Wiley, 33–97.

Kumbhakar, S. and C. Lovell (2000): *Stochastic Frontier Analysis*, Cambridge University Press.

Li, C., H. Li, and J. S. Racine (2014): "Cross-validated mixed datatype bandwidth selection for nonparametric cumulative distribution/survivor functions," Manuscript.

Li, Q. and J. Racine (2003): "Nonparametric estimation of distributions with categorical and continuous data," *Journal of Multivariate Analysis*, 86, 266–292.

Li, Q. and J. S. Racine (2008): "Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data," *Journal of Business and Economic Statistics*, 26, 423–434.

Martins-Filho, C. and F. Yao (2008): "A smooth nonparametric conditional quantile frontier estimator," *Journal of Econometrics*, 143, 317–333.

Meeusen, W. and J. van den Broeck (1977): "Efficiency estimation for Cobb-Douglas production functions with composed error," *International Economic Review*, 18, 435–444.

Park, B. U., L. Simar, and C. Weiner (2000): "The FDH estimator for productivity efficiency scores," *Econometric Theory*, 16, 855–877.

Parmeter, C. and S. Kumbhakar (2014): "Efficiency analysis: A primer on recent advances," *Foundations and Trends in Econometrics*, 7, 191–385.

Shephard, R. W. (1970): *Theory of Cost and Production Function*, Princeton, NJ: Princeton University Press.

Simar, L. and P. W. Wilson (2013): "Estimation and inference in non-parametric frontier models: Recent developments and perspectives," *Foundations and Trends(R) in Econometrics*, 5, 183–337.

Specification of frontier	order-m		α-quantile		α-quantile order-m	
	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
1. linear	1.500	1.365	0.056	0.028	0.006	0.001
2. step	1.498	1.395	0.010	0.003	0.010	0.002
3. quadratic	11.186	13.837	4.062	3.678	1.998	1.338

Table 3.1: Average root mean squared error (RMSE) for the nonsmooth order-m, α-quantile, and α-quantile order-m frontier estimators based on M=500 draws with sample sizes n=500 and n=1000. The RMSE for the order-m and the α-quantile estimators is the minimum for all possible choices of m and α, respectively. The RMSE for the α-quantile order-m estimator is obtained using, row-by-row, $(\alpha, m) = (0.25, 100), (0.25, 90), (0.25, 90), (0.25, 90), (0.2, 130), (0.25, 150)$.

Specification of frontier	α-quantile		α-quantile order-m	
	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
1. cubic	7.832	7.675	5.579	5.810
2. exponential	15.001	12.025	12.567	10.001

Table 3.2: Average root mean squared error (RMSE) for the smooth α-quantile and α-quantile order-m frontier estimators based on M=500 draws with sample sizes n=500 and n=1000. The RMSE for the α-quantile estimator is the minimum for all possible choices of α. The RMSE for the α-quantile order-m estimator is obtained using, row-by-row, $(\alpha, m) = (0.5, 50), (0.5, 50), (0.25, 90), (0.45, 70)$.

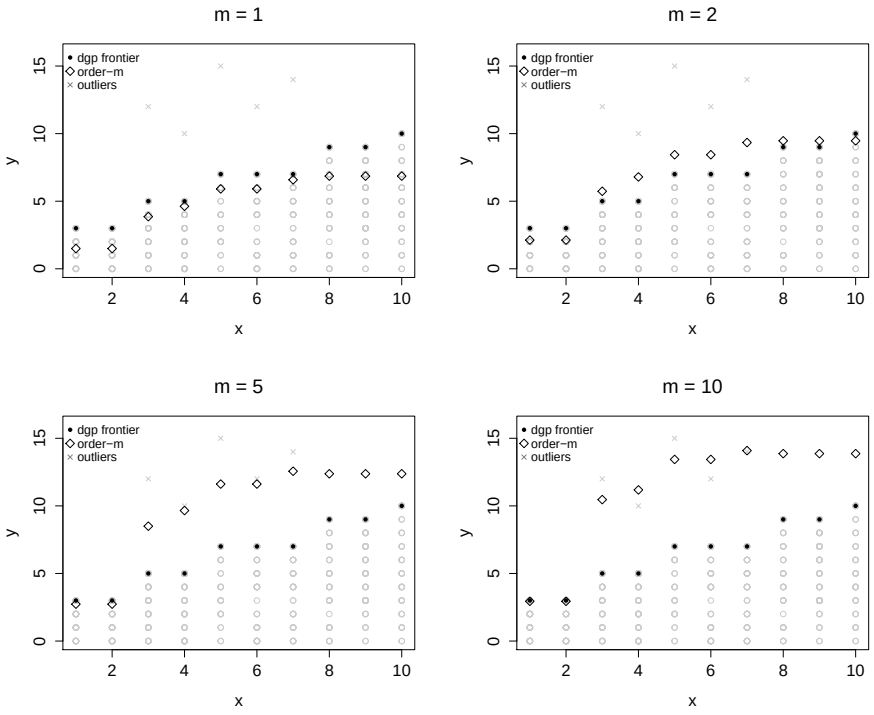

Figure 3.1: Order-m frontier estimates for simulated data with n = 500 (5 outliers).

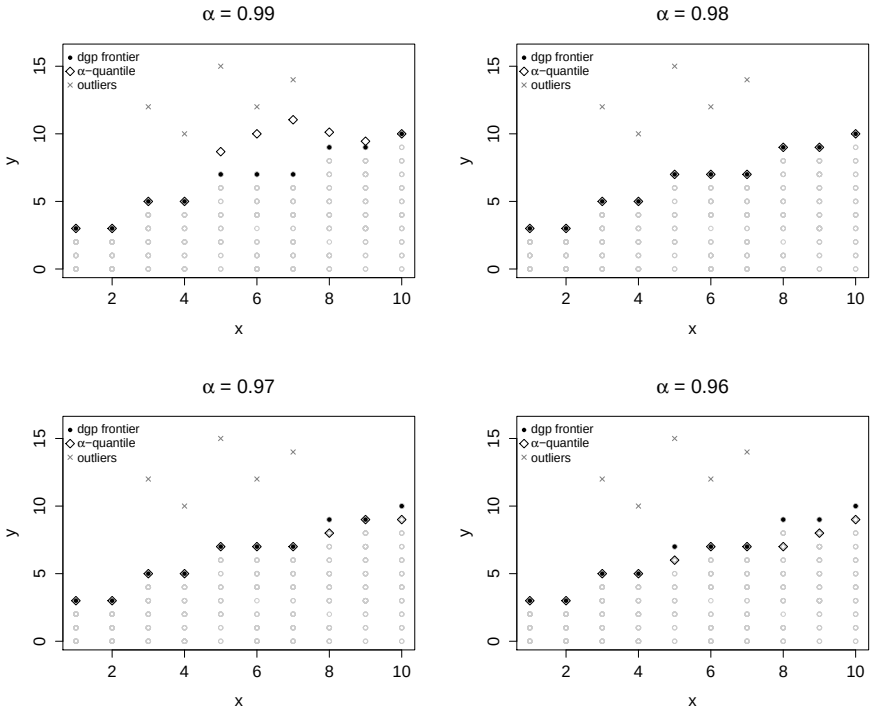

Figure 3.2: α-quantile frontier estimates for simulated data with n = 500 (5 outliers).

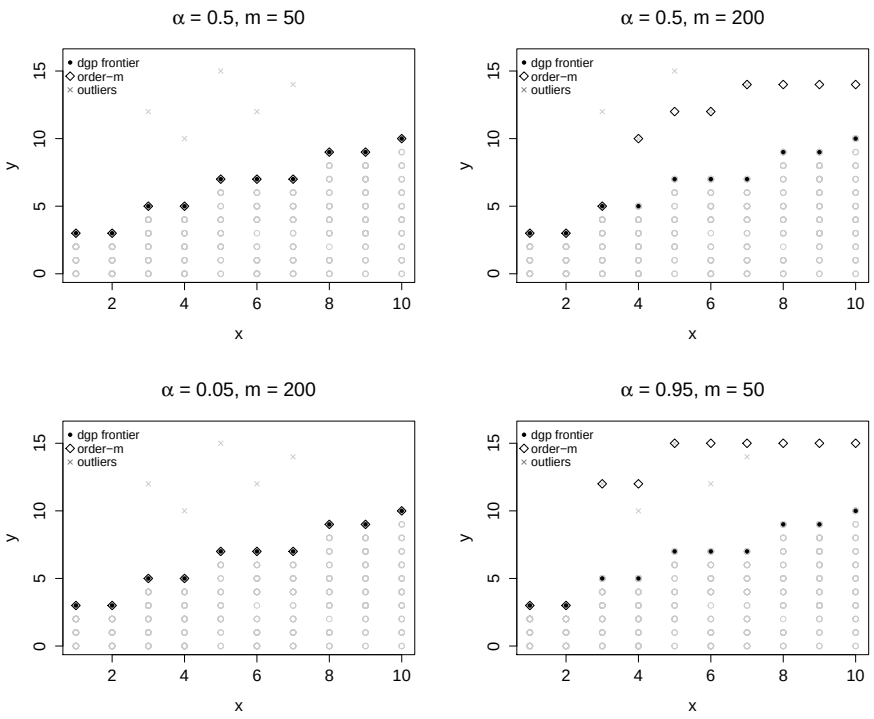

Figure 3.3: α-quantile order-m frontier estimates for simulated data with n = 500 (5 outliers).

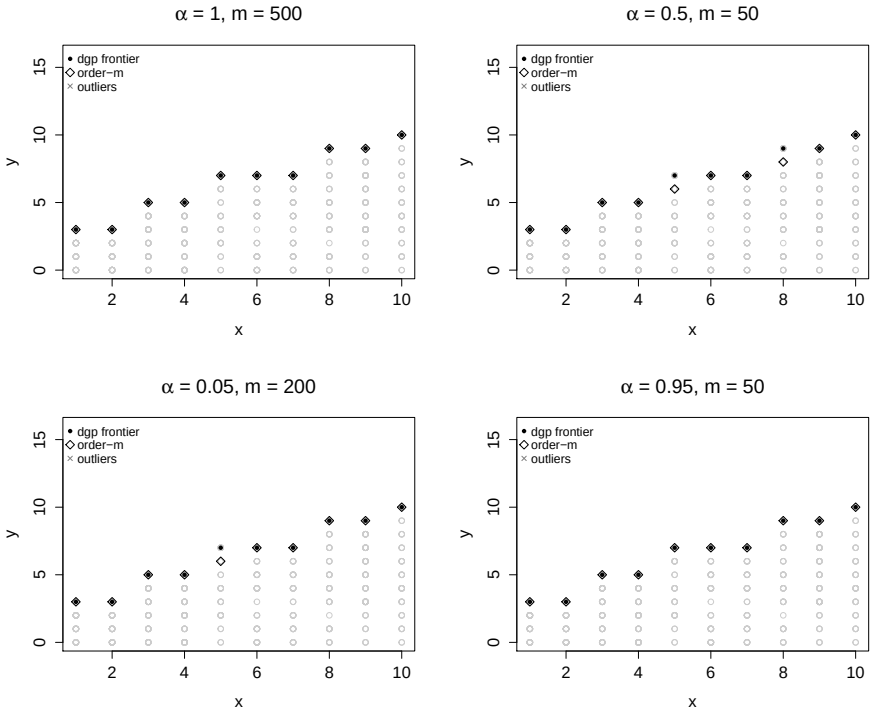

Figure 3.4: α-quantile order-m frontier estimates for simulated data with n = 500 (no outliers).

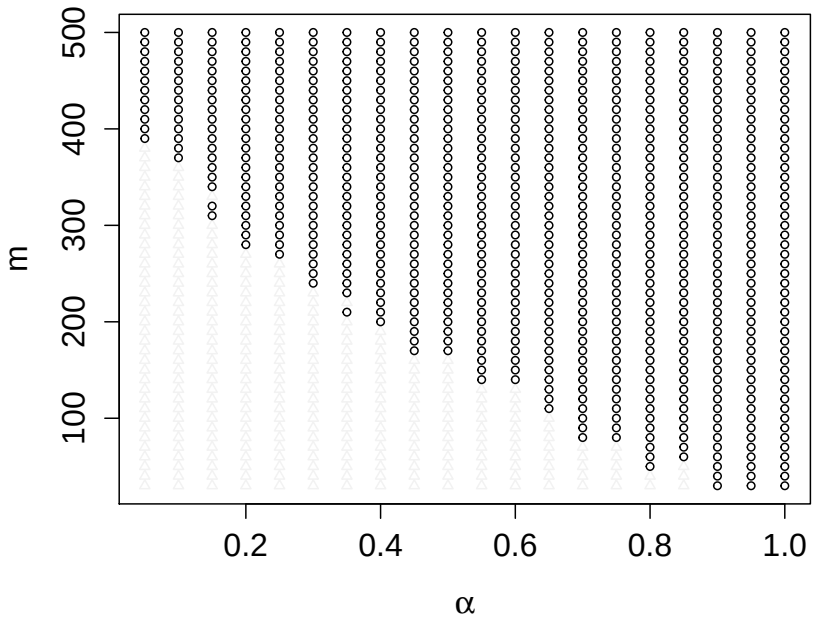

Figure 3.5: Clustering of α-quantile order-m frontier estimates.

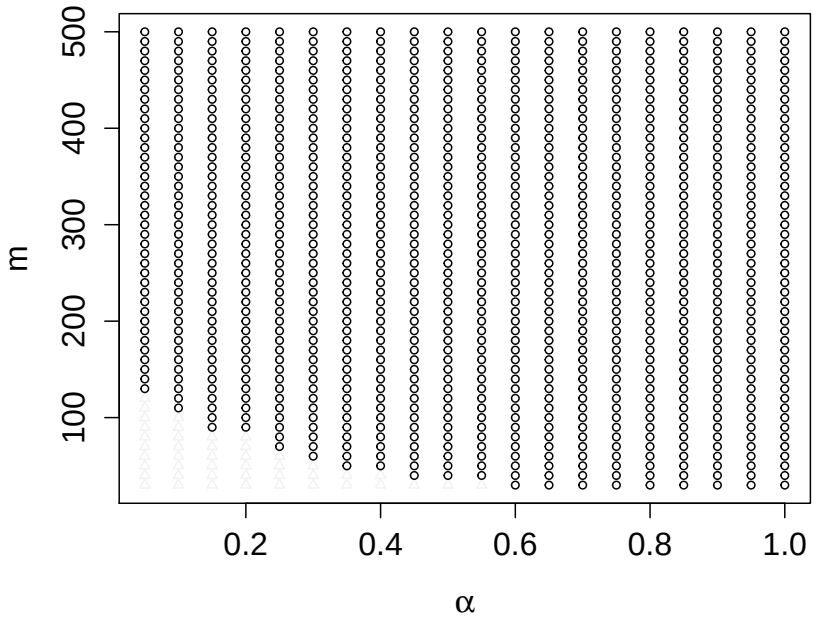

Figure 3.6: Clustering of α-quantile order-m frontier estimates (no outliers).

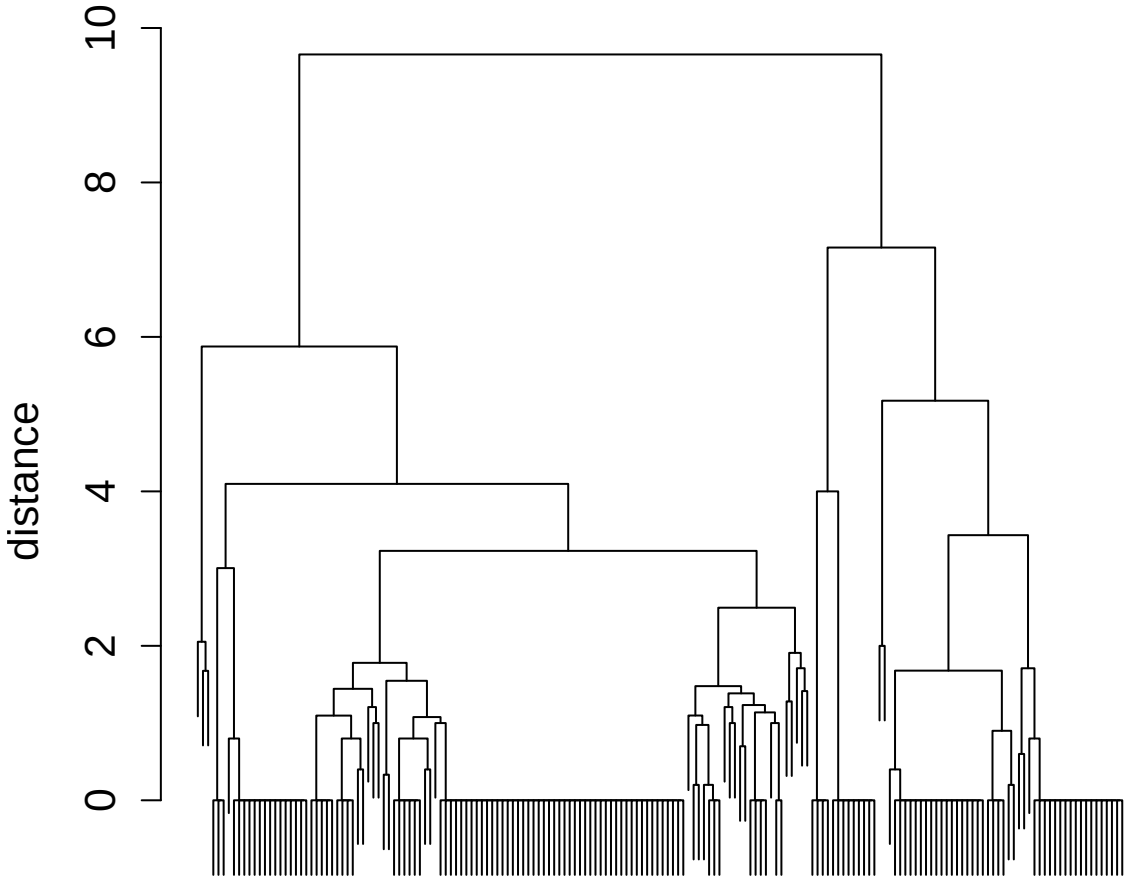

Figure 3.7: Dendrogram plot for hierarchical clustering procedure.

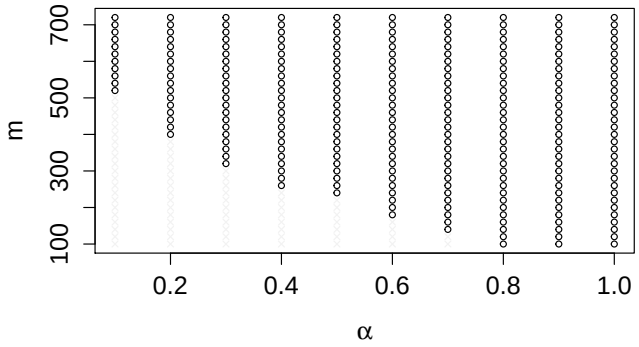

Figure 3.8: Clustering of α-quantile order-m patent count frontier estimates.

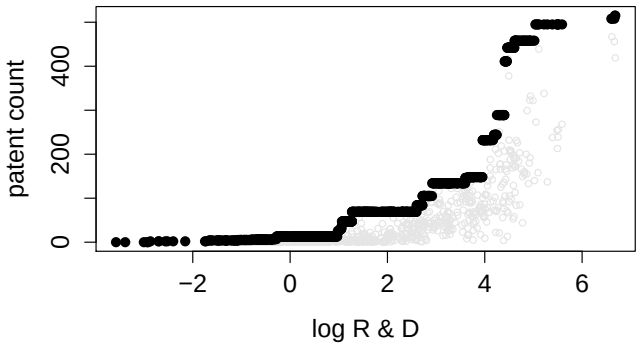

Figure 3.9: FDH patent count frontier ($\alpha = 1, m = n$).

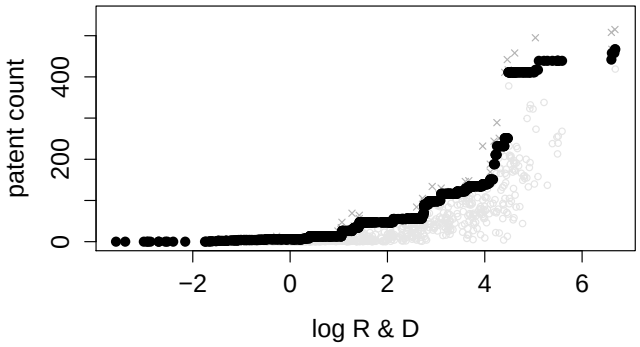

Figure 3.10: Nonsmooth α-quantile order-m patent count frontier ($\alpha = 0.3, m = 240$).

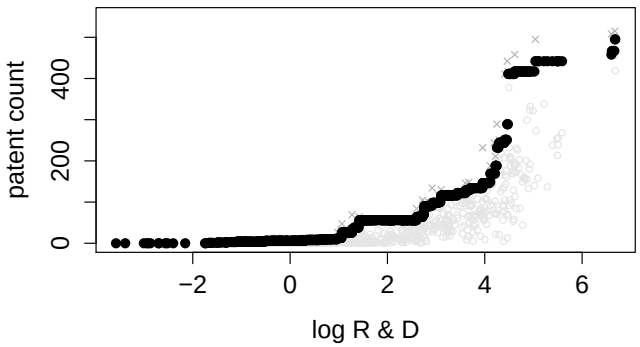

Figure 3.11: Smooth α-quantile order-m patent count frontier ($\alpha = 0.4, m = 200$).

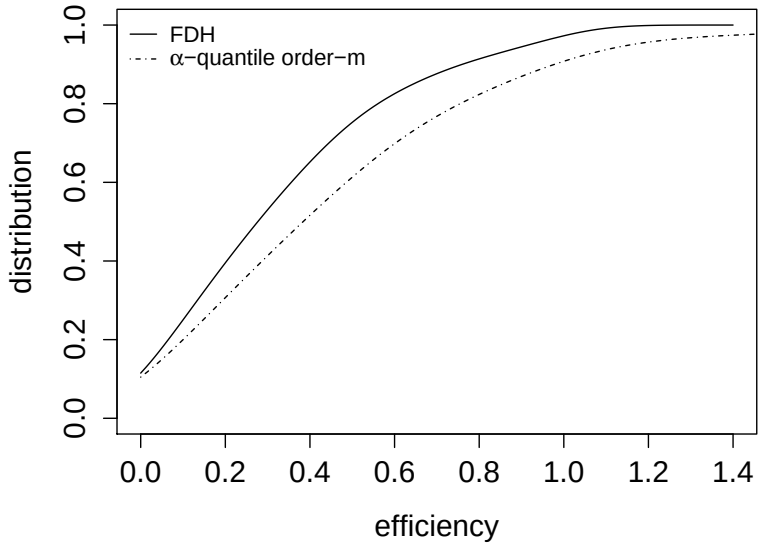

Figure 3.12: Distribution of patent count efficiency estimates (nonsmooth).

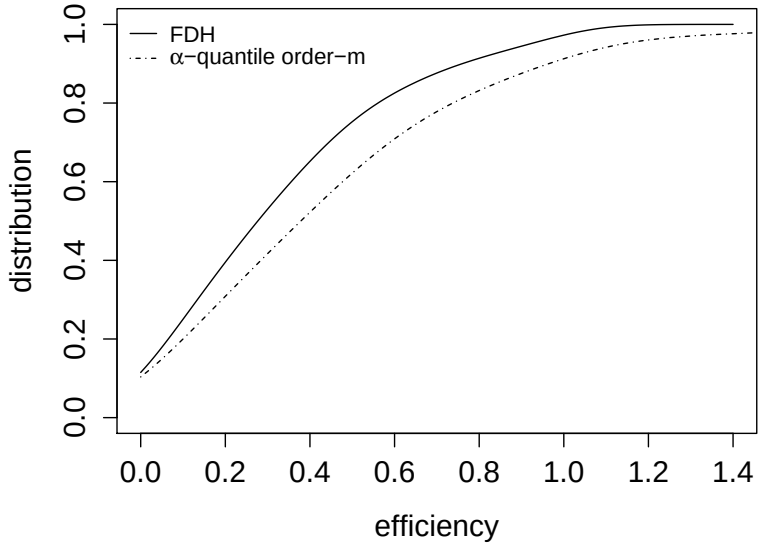

Figure 3.13: Distribution of patent count efficiency estimates (smooth).

Chapter 4

Nonparametric estimation of stochastic production frontier models for panel data

4.1 Introduction

Stochastic frontier modelling enjoys considerable popularity as a methodological framework for the analysis of firm-level production. Ever since the pioneering contributions of Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), the basic framework has been extended to cover a broad variety of model specifications in both cross-sectional and panel data settings. However, the literature has yet to capitalize on recent developments in the area of nonparametric conditional mean and gradient estimation against a panel data backdrop. This has been a lost opportunity, since stochastic frontier methods often need to be applied to datasets that have both a cross-sectional and a time series dimension, and nonparametric estimators offer a degree of flexibility that is simply not found in their parametric counterparts.

Hence, this chapter proposes a kernel-based estimation procedure for production frontier models that have both persistent and time-varying inefficiency components. It is shown that the parametric and nonparametric approaches yield substantially different estimates of i) the frontier itself, ii) factor elasticities, and iii) firm-level inefficiency.

The stochastic frontier literature is extremely vast and spans a period of nearly four decades. In the panel data context, Heshmati and Kumbhakar (1995), Greene (2005a,b), and Hardaker, Kumbhakar, and Lien (2014) outline a number of useful methods for the estimation of frontier models that include firm-specific intercepts and multi-dimensional measures of inefficiency comprising both persistent and time-varying elements. However, one of the downsides of the aforementioned approaches is their reliance on a parametric specification of the frontier function which, if incorrect, can give rise to imprecise estimates of both the frontier and the various components of firm-level inefficiency. Semiparametric and nonparametric estimators make it possible to sidestep the issue of model misspecification and hence, kernel-based approaches have sporadically appeared in the stochastic frontier literature. In particular, Fan, Li, and Weersink (1996), Parmeter and Racine (2012), and Martins-Filho and Yao (2015) apply kernel weighting methods to the estimation of a production frontier while maintaining parametric assumptions about the distribution of the inefficiency and stochastic error terms. However, both of these semiparametric estimation procedures are intended for cross-sectional data settings in which firm-level inefficiency is entirely time-invariant.

Fortunately, the econometric literature offers a variety of nonparametric regression techniques that are entirely suitable for stochastic frontier models

that involve panel data. For instance, Henderson and Ullah (2005), Su and Ullah (2007), and Martins-Filho and Yao (2009) propose kernel-based methods of estimating the types of error component models that are frequently used to model frontier functions and firm-level inefficiency. The first two of these approaches are based on a two-step locally-weighted generalized least squares procedure, while the third relies on a two-step transformation of the regression equation into one with errors that have a spherical parametric covariance structure. More recently, Ma, Racine, and Ullah (2015) have proposed a regression-spline random effects modelling framework that can be viewed as an alternative to two-step kernel-based methods. In the present chapter, we make use of Martins-Filho and Yao's (2009) procedure, although any of the aforementioned approaches would be appropriate in a panel data environment.

An ongoing challenge in the production frontier literature relates to the separate identification of time-varying inefficiency and stochastic noise in panel data settings. Identification is generally not possible without making parametric assumptions about the distribution of the former, but this can be problematic insofar as it introduces a risk of model misspecification. Horrace and Parmeter (2011) have built on previous work by Meister (2006) and outlined a semiparametric deconvolution procedure to recover the density of time-varying inefficiency, while Hall and Simar (2002) have proposed a kernel-based estimator for its mean; however, neither of these approaches can be used to measure each individual firm's distance from the production frontier. Kneip and Simar's (1996) nonparametric stochastic frontier estimator for panel data is somewhat promising, but it requires either i) an unrealistically-large number of observations for each cross-sectional unit or ii) an assumption that firm-level

inefficiency is entirely time-invariant. In contrast, the present chapter shows how it is possible to separately identify the persistent and time-varying components of inefficiency without making any parametric assumptions about the functional form of the production frontier or the distribution of the convoluted error term.

The chapter proceeds as follows: Section 4.2 provides an overview of both the baseline parametric stochastic frontier estimator and the nonparametric alternative that is being proposed. Section 5.3 discusses the firm-level dataset from the Colombian manufacturing sector that underlies the empirical analysis in Section 5.4, where evidence is provided that the parametric and nonparametric inefficiency estimates are characterized by a first-order stochastic dominance relationship. Section 4.5 presents the results of two specification tests that serve to scrutinize the distributional assumptions underlying the parametric frontier model. Section 4.6 comprises a Monte Carlo experiment that illustrates the adverse consequences of a misspecified frontier function in a finite sample environment, while Section 4.7 concludes.

4.2 Model specification and estimation strategy

This section provides an overview of the baseline parametric stochastic frontier model for panel data and a nonparametric alternative that can be estimated via kernel-based random effects regression. These two approaches are distinguished by their characterization of the frontier function and of firm-level inefficiency. The parametric framework relies on a set of functional form and distributional assumptions that the nonparametric framework renders unnecessary.

4.2.1 A parametric stochastic frontier model for panel data

Heshmati and Kumbhakar's (1995) parametric stochastic production frontier model is specified as:

$$y_{it} = \alpha_0 + \mathbf{x}'_{it}\boldsymbol{\beta} + e_{it} - v_{it} - u_i, \tag{4.1}$$

where y_{it} is the log of firm i's output in period t, $\mathbf{x}_{it}$ is a vector of inputs (capital, labour, etc.) with a corresponding vector $\boldsymbol{\beta}$ of factor elasticities, α_0 is an intercept that is common across all firms, and e_{it} is a i.i.d. $\mathcal{N}(0, \sigma_e)$ error term. The non-negative measure of technical inefficiency $v_{it} + u_i$ is divided into a time-varying ($v_{it} \geq 0$) and a persistent ($u_i \geq 0$) component; these components are independent both of one another and of e_{it}. Note that the non-negativity of v_{it} and u_i implies $\mathbb{E}(v_{it} + u_i|\mathbf{x}_{it}) \neq 0$ and hence, the production frontier $\alpha_0 + \mathbf{x}'_{it}\boldsymbol{\beta} \neq \mathbb{E}(y_{it}|\mathbf{x}_{it})$. Consequently, (4.1) cannot be estimated by following a standard panel data regression approach; instead, one can adopt the strategy that is outlined in Hardaker et al. (2014) and rewrite (4.1) as follows:

$$\begin{aligned} y_{it} &= \alpha_0 - \mathbb{E}(v_{it}) - \mathbb{E}(u_i) + \mathbf{x}'_{it}\boldsymbol{\beta} + e_{it} + \mathbb{E}(v_{it}) - v_{it} + \mathbb{E}(u_i) - u_i \\ &= \gamma_0 + \mathbf{x}'_{it}\boldsymbol{\beta} + \xi_{it} + \eta_i, \end{aligned} \tag{4.2}$$

where now, $\gamma_0 = \alpha_0 - \mathbb{E}(v_{it}) - \mathbb{E}(u_i)$, $\xi_{it} = e_{it} + \mathbb{E}(v_{it}) - v_{it}$, and $\eta_i = \mathbb{E}(u_i) - u_i$. This alternative specification of the model ensures that $\mathbb{E}(\xi_{it}|\mathbf{x}_{it}) = 0$ and $\mathbb{E}(\eta_i|\mathbf{x}_{it}) = 0$, which clears the way for estimation of the conditional mean

$\mathbb{E}(y_{it}|\mathbf{x}_{it}) = \gamma_0 + \mathbf{x}'_{it}\boldsymbol{\beta}$ by means of a random effects regression.[1]

Given that the random effects regression yields predicted values of $\eta_i = \mathbb{E}(u_i) - u_i$, it is possible to estimate the persistent component u_i of firm-level inefficiency by adopting the method that is outlined in Heshmati and Kumbhakar (1995) and Hardaker et al. (2014):

$$\hat{u}_i = \max(\hat{\eta}_i) - \hat{\eta}_i. \tag{4.3}$$

Next, in order to separate v_{it} and e_{it}, it is standard practice in the stochastic frontier literature to assume that v_{it} is characterized by either a half-normal[2] or an exponential distribution. In this setting, the time-varying component of firm-level inefficiency can be estimated via maximum likelihood. Given that $e_{it} \sim \mathcal{N}(0, \sigma_e)$, the density function that underlies $\xi'_{it} = e_{it} - v_{it} = \xi_{it} - \mathbb{E}(v_{it})$ is specified as either normal-half-normal or normal-exponential:

$$f(\xi'_{it}) = \frac{2}{\sigma}\phi\left(\frac{\xi'_{it}}{\sigma}\right)\Phi\left(-\frac{\xi'_{it}\lambda}{\sigma}\right) \tag{4.4}$$

$$f(\xi'_{it}) = \frac{1}{\sigma_v}\Phi\left(-\frac{\xi'_{it}}{\sigma_e} - \frac{\sigma_e}{\sigma_v}\right)\cdot e^{\frac{\xi'_{it}}{\sigma_v} + \frac{\sigma_e^2}{2\sigma_v^2}}, \tag{4.5}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal probability density and cumulative distribution functions, respectively, $\sigma = \sqrt{\sigma_v^2 + \sigma_e^2}$, and $\lambda = \frac{\sigma_v}{\sigma_e}$. Parmeter and Kumbhakar (2014) show that the log-likelihood functions cor-

[1] The random effects regression framework assumes that $\mathbf{x}_{it}$ is independent of v_{it} and u_i, that is, productive inefficiency does not influence the input decisions of firms. Alternatively, one can adopt the fixed effects specification proposed by Schmidt and Sickles (1984) and summarized in Parmeter and Kumbhakar (2014), which does not require independence of $\mathbf{x}_{it}$, v_{it}, and u_i.

[2] A half-normal distribution $\mathcal{N}^+(0, \sigma)$ is simply a $\mathcal{N}(0, \sigma)$ distribution with restricted domain $[0, \infty)$.

responding to (4.4) and (4.5) are expressed as:

$$\ln \mathcal{L} = -NT \ln \sigma + \sum_{i=1}^{N} \sum_{t=1}^{T} \ln \Phi \left(-\frac{\xi'_{it} \lambda}{\sigma} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \xi'^2_{it} \tag{4.6}$$

$$\ln \mathcal{L} = -NT \ln \sigma_v + NT \left(\frac{\sigma_e^2}{2\sigma_v^2} \right) + \sum_{i=1}^{N} \sum_{t=1}^{T} \ln \Phi \left(-\frac{\xi'_{it}}{\sigma_e} - \frac{\sigma_e}{\sigma_v} \right) + \frac{1}{\sigma_v} \sum_{i=1}^{N} \sum_{t=1}^{T} \xi'_{it}. \tag{4.7}$$

Note that one can substitute $\xi'_{it} = y_{it} - \hat{\gamma}_0 - \mathbf{x}'_{it}\hat{\boldsymbol{\beta}} - \hat{\eta}_i - \mathbb{E}(v_{it})$ into (4.6) and (4.7), where $\hat{\gamma}_0$ and $\hat{\boldsymbol{\beta}}$ are the random effects regression coefficient estimates from (4.2). The unknown mean $\mathbb{E}(v_{it}) = \sqrt{\frac{2}{\pi}} \cdot \sigma_v$ when the assumed distribution of v_{it} is half-normal and $\mathbb{E}(v_{it}) = \sigma_v$ when it is exponential. Fan et al. (1996) also show how a bit of algebra makes it possible to express the normal-half-normal likelihood function in terms of the single unknown parameter λ, which results in a simplified optimization problem. Once the parameters λ, σ, σ_e, σ_v, and $\mathbb{E}(v_{it})$ have been estimated, the production frontier is obtained by adding $\widehat{\mathbb{E}(u_i)}$ and $\widehat{\mathbb{E}(v_{it})}$ to the fitted values $\hat{\gamma}_0 + \mathbf{x}'_{it}\hat{\boldsymbol{\beta}}$ from the random effects regression at the outset. Finally, depending on which of the half-normal or exponential distributional assumptions have been adopted, the method of either Jondrow, Lovell, Materov, and Schmidt (1982) or Kumbhakar and Lovell (2000) can be used to compute the time-varying component of firm i's period-t inefficiency $\hat{v}_{it}$.

4.2.2 A nonparametric stochastic frontier model for panel data

Consider a setting in which neither the functional form of the production frontier nor the distribution of firm-level inefficiency is subject to any parametric

assumptions. In particular, suppose the model is now expressed as:

$$y_{it} = g_t(\mathbf{x}_{it}) + e_{it} - u_i, \tag{4.8}$$

where $g_t(\cdot) \equiv g(\cdot, t)$ denotes the frontier function, $u_i \geq 0$ denotes persistent inefficiency, and e_{it} is a stochastic noise term. In the baseline parametric model for panel data that was discussed in the previous subsection, the maximum output quantity that can be achieved with the input vector $\mathbf{x}_{it}$ does not change over time, but this seems at odds with intuitive conceptualizations of technological progress and productivity growth. Instead, a more appropriate assumption might be that a firm's output lies somewhere on or below a production frontier whose location and shape are nevertheless evolving at each $t = 1, ..., T$. The varying component of inefficiency v_{it} that was discussed in Section 4.2.1 has been dropped from the analysis and, instead, the frontier function includes a time subscript $g_t(\cdot)$. Given that the (unknown) distribution of u_i has non-negative support, it cannot be assumed that $\mathbb{E}(u_i|\mathbf{x}_{it}) = 0$ and hence, (4.8) must be rewritten as:

$$y_{it} = h_t(\mathbf{x}_{it}) + e_{it} + \eta_i, \tag{4.9}$$

where $h_t(\mathbf{x}_{it}) = g_t(\mathbf{x}_{it}) - \mathbb{E}(u_i)$ and $\eta_i = \mathbb{E}(u_i) - u_i$.

The conditional mean function $h_t(\mathbf{x}_{it}) = g_t(\mathbf{x}_{it}) - \mathbb{E}(u_i)$ and the $q \times 1$ vector of factor elasticities $\nabla h_t(\mathbf{x}_{it}) = \left(\frac{\partial h_t}{\partial x_1}(\mathbf{x}_{it}), ..., \frac{\partial h_t}{\partial x_q}(\mathbf{x}_{it})\right) = \left(\frac{\partial g_t}{\partial x_1}(\mathbf{x}_{it}), ..., \frac{\partial g_t}{\partial x_q}(\mathbf{x}_{it})\right)$ are estimated using the two-step, kernel-based nonparametric random effects estimator of Martins-Filho and Yao (2009).[3] In short, the estimation procedure

[3] Alternatively, one may use the kernel estimators of Henderson and Ullah (2005) or Su

can be delineated as follows:

1. Estimate (4.9) by means of a pooled local quadratic[4] regression

2. Use the residuals from step 1 to compute the variance estimates $\hat{\sigma}_e^2$ and $\hat{\sigma}_\eta^2$, just as one would do under the parametric random effects framework

3. Use $\hat{\sigma}_e^2$ and $\hat{\sigma}_\eta^2$ to transform the regression equation into one with errors that have a spherical covariance structure

4. Estimate the transformed equation using a local quadratic regression

The inclusion of the discrete time variable t in the model implies that the nonparametric random effects estimator must admit a mixture of continuous and discrete predictors. To this end, Li and Racine's (2004) generalized product kernel is used to construct a local weighting function, which is then incorporated into Martins-Filho and Yao's (2009) proposed framework. A more thorough summary of the procedure that is used to estimate the conditional mean function $\mathbb{E}_t\left(y_{it}|\mathbf{x}_{it}\right) = h_t\left(\mathbf{x}_{it}\right)$ can be found in the appendix. In practice, the method is implemented using the generalized local polynomial regression and kernel summation functions in the 'crs' and 'np' packages, respectively, in the R statistical computing environment (Hayfield and Racine, 2008).

Note that one of the primary advantages of the nonparametric random effects estimator is that it is fully compatible with heterogeneous firm-level factor elasticities. Given that the estimates of $\nabla h_t\left(\mathbf{x}_{it}\right) = \nabla g_t\left(\mathbf{x}_{it}\right) = \left(\frac{\partial g_t}{\partial x_1}\left(\mathbf{x}_{it}\right), ..., \frac{\partial g_t}{\partial x_q}\left(\mathbf{x}_{it}\right)\right)$

and Ullah (2007), or the spline-based approach that is delineated in Ma et al. (2015). Under a fixed effects specification, the nonparametric estimator of Lee and Robinson (2015) might be an option, although this particular framework is not considered here.

[4]The local quadratic specification is favoured over the local linear specification because it gives rise to better estimates of the first derivative of the conditional mean function.

are functions of the firm-specific input vector $\mathbf{x}_{it}$ and of time, the method allows for both intra-industry and intertemporal variation in the elasticity of output with respect to $x_1, ..., x_q$. This ensures a far greater degree of flexibility than the conventional parametric random effects framework, which tends to rely on the rather rigid assumption of constant elasticities. An issue raised in Kealey (2015) is that even though the log-linear Cobb-Douglas specification of a firm-level production function has proven popular in applied research[5], it is nonetheless incompatible with an elementary theoretical result that equates factor elasticities and their corresponding input expenditure shares. Meanwhile, one can expect the kernel-based estimator to yield a more realistic degree of co-movement between these expenditure shares and the partial derivatives in $\widehat{\nabla g_t}(\mathbf{x}_{it})$.

The nonparametric random effects regression yields the predicted values $\hat{h}_t(\mathbf{x}_{it})$ and $\hat{\eta}_i$. Subsequent application of the formula in (4.3) yields $\hat{u}_i$, $\widehat{\mathbb{E}(u_i)}$, and $\hat{g}_t(\mathbf{x}_{it}) = \hat{h}_t(\mathbf{x}_{it}) + \widehat{\mathbb{E}(u_i)}$, which can be applied to the computation of the proposed time-varying measure of firm-level inefficiency. This involves two simple steps:

1. Define $\hat{g}_{t^*}(\mathbf{x}) = \max_{t=1,...,T} \hat{g}_t(\mathbf{x})$ for any input vector $\mathbf{x}$

2. Firm i's period-t inefficiency is given by $\hat{\gamma}_{it} = \hat{g}_{t^*}(\mathbf{x}_{it}) - y_{it} + \hat{e}_{it}$

Importantly, $\hat{\gamma}_{it}$ is obtained without making any parametric assumptions about its underlying density function. Moreover, it can be decomposed in a manner

[5]See, for instance, Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg, Caves, and Frazer (2006), and Wooldridge (2009).

that elucidates the intuition behind the proposed approach:

$$\begin{aligned}\hat{\gamma}_{it} &= \hat{g}_{t^*}(\mathbf{x}_{it}) - y_{it} + \hat{e}_{it} \\ &= \hat{g}_t(\mathbf{x}_{it}) - y_{it} + \hat{e}_{it} - (\hat{g}_t(\mathbf{x}_{it}) - \hat{g}_{t^*}(\mathbf{x}_{it})) \\ &= \hat{u}_i + (\hat{g}_{t^*}(\mathbf{x}_{it}) - \hat{g}_t(\mathbf{x}_{it}))\,.\end{aligned} \tag{4.10}$$

Thus, $\hat{\gamma}_{it}$ comprises firm i's distance from the frontier in the current period ($\hat{u}_i$) in addition to the distance between the frontier functions $\hat{g}_{t^*}$ and $\hat{g}_t$ when they are both evaluated at $\mathbf{x}_{it}$.[6] If the production frontier is expanding and/or contracting over time (which, it is argued here, ought to be interpreted as a form of productivity growth), then this will result in changing values of $\hat{\gamma}_{it}$ when $t = 1, ..., T$. We now turn to a comparison of this new measure of firm-level inefficiency with the one that was outlined in Section 4.2.1.

4.3 Data

The balanced panel dataset that underlies the empirical analysis in Section 5.4 of this chapter comprises 12,749 observations on Colombian manufacturing plants over the period 1981-1991. Four broadly defined industries are represented in the sample: food processing, textiles and apparel, furniture and finished wood products, and fabricated metal. The data were originally collected in a country-wide industrial census whose coverage extended to all manufacturers with 10 or more employees. This Colombian dataset has appeared previously in a number of studies that are concerned with the estimation of firm-level productivity (Roberts and Tybout, 1997; Fernandes, 2007; Gandhi,

[6]Note that $\hat{g}_{t^*}(\mathbf{x}_{it}) - \hat{g}_t(\mathbf{x}_{it}) \geq 0$ by construction and hence, $\hat{\gamma}_{it} \geq \hat{u}_i$ for $t = 1, ..., T$.

Navarro, and Rivers, 2016).

The production models in (4.1) and (4.8) are estimated under both a gross output and a value-added specification. In the context of the former, y_{it} denotes the natural log of firm i's gross output in period t, while the vector $\mathbf{x}_{it}$ is composed of the natural log of five inputs, namely capital, unskilled labour, skilled labour, raw materials, and energy. The capital stock variable is constructed by taking the sum of land, buildings/structures, machinery, transportation equipment, and office equipment, whose respective values are computed using the perpetual inventory method and 3-digit industry-level depreciation data found in Pombo (1999). The unskilled and skilled labour inputs are expressed in terms of the number of workers employed by a plant in a particular year, while the raw materials and energy consumption variables (i.e., the intermediate inputs in the gross-output specification of the production function) are measured in thousands of Colombian pesos. In accordance with the convention that is followed in much of the productivity literature, the model is estimated using logarithmic transforms of each of the entries in $\mathbf{x}_{it}$. In addition, prior to their log transformation, the gross output, capital stock, and intermediate input variables are all deflated by an industry-year-specific index[7] so that the common unit of measurement across the entire panel is thousands of pesos at 1991 price levels.

Under the alternative value-added specification of the models in (4.1) and (4.8), y_{it} denotes the log of the difference between a firm's gross output and its consumption of raw materials and energy. Meanwhile, the input vector $\mathbf{x}_{it}$ is composed of the capital stock and the skilled/unskilled labour variables.

[7]The price index is computed as the ratio of the nominal to the real value of production for any given firm in a particular industry and year.

The value-added production function might be preferred over the gross output version of the model if there is reason to believe that the efficiency of a firm (either persistent or time-varying or both) tends to influence its intermediate input decisions; the interplay between firm productivity on the one hand, and chosen variable input quantities on the other, and the ensuing risk that factor elasticity estimates will suffer from "transmission bias", has been the subject of much discussion in the econometrics and industrial organization literature (Marschak and Andrews, 1944; Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2006).

Figure 4.2 illustrates the distribution of firm-level factor shares in the food processing, textile and apparel, furniture and finished wood products, and fabricated metal industries. The four boxplots respectively correspond to the capital stock, skilled wages and benefits, unskilled wages and benefits, and intermediate inputs, and each of these is expressed as a share of firms' gross output in each year between 1981 and 1991. It is clear that factor shares exhibit a considerable amount of heterogeneity across firms and over time. Firm-level capital-output ratios are quite dispersed and are generally increasing over the 11 year period, whereas the fraction of output being allocated to wage and benefit payments to unskilled workers appears to be in decline. This raises doubts about the specification of the frontier model in Section 4.2.1, where corresponding factor elasticities are treated as fixed parameters.

4.4 Results

The results that are presented in this section pertain to the parametric and nonparametric estimates of factor elasticities and firm-level inefficiency (i.e.

each firm's measured distance from the production frontier). The supporting figures that are referenced in what follows can be found in Appendix B.

4.4.1 Factor elasticity estimates

As mentioned earlier in Section 4.2.2, an important difference between the parametric and nonparametric frontier models is that the former treats factor elasticities as fixed while the latter allows them to vary across firms and over time. Figures 4.3 and 4.4 provide a clear illustration of this contrast. Boxplots of the input elasticity estimates obtained using the kernel-based random effects estimator suggest that there is a considerable amount of intra-industry and intertemporal heterogeneity in the partial derivative of gross output with respect to capital, skilled labour, unskilled labour, energy, and raw materials. This is consistent with the substantial variation in Colombian manufacturing plants' input expenditure shares that was noted in Section 5.3. Boxplots for the value-added model where capital and labour are the only inputs are also provided in figure 4.5, and the results are qualitatively similar to those obtained under the gross output specification. Table 4.1 presents Spearman correlations of the nonparametric factor elasticity estimates and their corresponding expenditure shares. In nearly every instance, the sign and magnitude of the correlation coefficients are consistent with what standard theories of firm behaviour would tend to predict. This is particularly true for the two largest industries in the sample, namely food processing and textiles/apparel manufacturing.

In all four industries, it appears that plant-level production technology becomes less and less dependent on unskilled labour over the 11-year period 1981-1991, which has also been discussed in Kealey (2015) in the context of

a varying-coefficient estimator for a Cobb-Douglas production function. Furthermore, the food processing, textiles/apparel, furniture/finished wood products, and fabricated metal industries appear to become more capital-intensive toward the late 1980s and early 1990s. Note that this is entirely consistent with the time trends that were observed in the input expenditure boxplots when basic features of the plant-level dataset were discussed in Section 5.3. For the sake of comparison, parametric elasticity estimates from the model in (4.1) have been superimposed in dashed lines onto the boxplots in figures 4.3 through 4.5. Clearly, modelling the production frontier as a simple linear function leads one to either underestimate or overestimate the factor elasticities of most of the firms in the sample - a complication that is for the most part avoided when one follows the nonparametric estimation procedure that is being proposed in this chapter.

4.4.2 Firm-level inefficiency estimates

To begin, we would like to consider whether the parametric and nonparametric frontier methods yield broadly similar estimates of firm-level inefficiency. That is, even though the two approaches model the frontier function and factor elasticities in very different ways, it is entirely possible that this makes little difference for the measurement of firms' distance from the production frontier. We make use of Li, Maasoumi, and Racine's (2009) nonparametric test for equality of densities to determine whether the two samples of inefficiency estimates were drawn from the same hypothetical distribution. To begin, let $\{\hat{v}_{it} + \hat{u}_i\}_{i,t=1}^{NT}$ denote the parametric estimates from the stochastic frontier model in Section 4.2.1 and let $\{\hat{\gamma}_{it}\}_{i,t=1}^{NT}$ denote the nonparamet-

ric inefficiency estimates from equation (4.10) in Section 4.2.2. If $f(\cdot)$ and $g(\cdot)$ are the density functions that underlie $\hat{v}_{it} + \hat{u}_i$ and $\hat{\gamma}_{it}$, respectively, the object of interest is the integrated squared difference $\int [f(z) - g(z)]^2 dz = \int [f(z) dF(z) + g(z) dG(z) - f(z) dG(z) - g(z) dF(z)]$, where $F(\cdot)$ and $G(\cdot)$ are the cumulative distribution functions for $\hat{v}_{it} + \hat{u}_i$ and $\hat{\gamma}_{it}$, respectively. This integrated squared difference forms the basis of Li et al.'s (2009) test statistic:

$$\begin{aligned} I_{NT} = & \frac{1}{NT(NT-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{i' \neq i}^{N} \sum_{t' \neq t}^{T} k\left(\frac{\hat{v}_{it} + \hat{u}_i - \hat{v}_{i't'} - \hat{u}'_i}{h}\right) \\ & + \frac{1}{NT(NT-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{i' \neq i}^{N} \sum_{t' \neq t}^{T} k\left(\frac{\hat{\gamma}_{it} - \hat{\gamma}_{i't'}}{h}\right) \\ & - \frac{2}{(NT)^2} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{i'=1}^{N} \sum_{t'=1}^{T} k\left(\frac{\hat{v}_{it} + \hat{u}_i - \hat{\gamma}_{i't'}}{h}\right), \end{aligned} \tag{4.11}$$

where $k(\cdot)$ denotes a univariate Gaussian kernel function and h is a smoothing parameter. The null distribution of I_{NT} must be approximated using a bootstrap procedure. In practice, the nonparametric specification test can be carried out using the npdeneqtest() function in Hayfield and Racine's (2008) 'np' package in the R statistical computing environment.

Tables 4.2 and 4.3 show that the null hypothesis of equality of densities is unequivocally rejected in each of the 16 instances where the test is performed. The computed p-values are all equal to zero up to several decimal places, which suggests that the parametric and nonparametric frontier models give rise to substantially different estimates of firm-level inefficiency. While this outcome might seem intuitive, it is in fact a bit surprising. Regardless of which of the two methods is adopted, estimation of the conditional mean of the output variable $\mathbb{E}(y_{it}|\mathbf{x}_{it})$ and the persistent component of inefficiency u_i is carried out by means of a random effects regression. Thus, even though they rely on differing

assumptions about the functional form of $\mathbb{E}(y_{it}|\mathbf{x}_{it})$, the approaches that are delineated in Sections 4.2.1 and 4.2.2 need not produce dissimilar estimates of u_i. However, this appears to be what has occurred. Consider, for example, the gross output version of the parametric frontier model. Estimation of the time-varying component of inefficiency v_{it} via maximization of the log-likelihood function in (4.6) results in $\hat{\sigma}_v = 0$ in all four industries and hence, the measure of inefficiency ends up not having a time-varying component at all (i.e., the normal-half-normal distribution characterizing $e_{it} - v_{it}$ collapses to the $\mathcal{N}(0, \sigma_e)$ distribution of the stochastic noise term). That is, the parametric inefficiency estimate is simply given by $\hat{u}_i$. Meanwhile, in the nonparametric setting, decomposition of $\hat{\gamma}_{it} = \hat{u}_i + (\hat{g}_{t^*}(\mathbf{x}_{it}) - \hat{g}_t(\mathbf{x}_{it}))$ in (4.10) into its persistent and time-varying elements reveals that $\hat{u}_i$ tends to account for 90 percent or more of a firm's measured distance from the production frontier. Hence one might expect the parametric and nonparametric estimates of inefficiency under the value-added specification of the model to be rather similar, but this is not what is observed in the empirical example that involves the Colombian manufacturing data.

Figures 4.6 and 4.7 comprise plots of the parametric and nonparametric inefficiency estimates' empirical distributions. Interestingly, there appear to be considerable differences in the magnitude of the estimates that arise from the two competing methods. In particular, in all four industries, the distance that separates firms' output from the production frontier tends to be larger under the parametric framework than under the nonparametric framework. From a practical point of view, this result has important implications insofar as it suggests that firms' productive efficiency may be either understated

or overstated depending on the model specification that has been chosen. In fact, casual observation of the empirical distribution plots in figures 4.6 and 4.7 raises questions about whether there might even exist a stochastic dominance relationship between the two classes of firm-level inefficiency estimates. This proposition can be examined by means of the statistical test for first-order stochastic dominance that was originally proposed by Klecan, McFadden, and McFadden (1991) and subsequently extended by Linton, Maasoumi, and Whang (2005).

4.4.3 A test for first-order stochastic dominance

To begin, let $F(\cdot)$ and $G(\cdot)$ denote the distribution functions that underlie the parametric and nonparametric estimates of inefficiency, respectively. We would like to determine whether the parametric inefficiency estimates first-order stochastically dominate the nonparametric ones, that is, whether $F(z) \leq G(z)$ for all nonnegative values of z. Given that the aforementioned CDFs do not have a known functional form, we make use of the empirical distribution functions that are defined as follows:

$$\hat{F}(z) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left(\hat{v}_{it} + \hat{u}_i \leq z\right) \quad ; \quad \hat{G}(z) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} 1\left(\hat{\gamma}_{it} \leq z\right), \tag{4.12}$$

where (4.12) uses the same notation as the frontier models in (4.1) and (4.10) where the parametric and nonparametric estimates of inefficiency are given by $\hat{v}_{it} + \hat{u}_i$ and $\hat{\gamma}_{it}$, respectively. To test the null hypothesis that there exists a first-order stochastic dominance (FOSD) relationship between the two groups

of inefficiency estimates, we make use of Klecan et al.'s (1991) test statistic:

$$D = \min\left\{\max_z[\hat{F}(z) - \hat{G}(z)], \max_z[\hat{G}(z) - \hat{F}(z)]\right\}, \tag{4.13}$$

where in practice, the maximum in (4.13) can be taken over some interval in $\mathbb{R}^+$ that includes the support of both $\hat{F}(\cdot)$ and $\hat{G}(\cdot)$. The null and alternative hypotheses are stated as:

$$\begin{aligned} H_0 &: D \leq 0 \\ H_a &: D > 0 \end{aligned} \tag{4.14}$$

Linton et al. (2005) propose a bootstrap procedure for obtaining a critical value D^* such that H_0 is rejected if $D > D^*$. For a given number of bootstrap iterations B, let $n = N - B + 1$ be the size of the subsample (without replacement) of firms that is used in each $b = 1, ..., B$. If the manufacturing firms in the sample are indexed by j, then for each $b = 1, ..., B$, we re-estimate firm-level inefficiency using observations $j = b, ..., n + b - 1$, define $F^{(b)}(\cdot)$ and $G^{(b)}(\cdot)$ as in (4.12), and compute $D^{(b)}$ as in (4.13). The critical value D^* is given by the $(1-\alpha)$-quantile of the $\left\{D^{(b)}\right\}_{b=1}^{B}$. In this chapter, we compute the sampling distribution of the test statistic using $B = 199$ bootstrap iterations for the food processing, textiles/apparel, and fabricated metal industries, and $B = 79$ iterations for the furniture/finished wood products industry due to sample size limitations.[8]

[8]Resampling occurs at the cross-sectional level, and given that $n = N - B + 1$, there is a tradeoff between the number of bootstrap iterations and the size of the subsample for each $b = 1, ..., B$. In the finished wood products industry, there are only $N = 104$ unique firms in the data with $T = 11$ observations each. Using $B = 79$ implies that each subsample comprises $n = (104-79+1)\cdot 11 = 286$ observations. If we opted for more bootstrap iterations, i.e., $B = 99$ instead of $B = 79$, that would only leave us with $n = (104 - 99 + 1) \cdot 11 = 66$,

Tables 4.4 and 4.5 present one of this chapter's central findings, namely that the parametric and nonparametric inefficiency estimates exhibit a FOSD relationship in 11 of the 16 instances where Klecan et al.'s (1991) test is performed. Furthermore, this result is not very sensitive to the distributional assumptions (i.e. normal-half-normal vs. normal-exponential) that underlie the time-varying components of the parametric frontier model. Given that the null hypothesis of the test is $F(z) \leq G(z)$ for all nonnegative values of z, where $F(\cdot)$ and $G(\cdot)$ respectively denote the CDFs of the parametric and nonparametric inefficiency estimates, a p-value that is substantially greater than zero should serve as evidence of FOSD. In the 11 different cases where we fail to reject the null hypothesis in (4.14), the p-values appearing in parentheses in tables 4.4 and 4.5 range from 0.19 to 1, and are frequently greater than 0.5; hence, there is reason to believe that Heshmati and Kumbhakar's (1995) parametric frontier model tends to generate inflated estimates of firm-level inefficiency when the nonparametric framework is used as a benchmark. While the results of the formal test for FOSD are indeed important, they are not particularly surprising in the light of the empirical CDF plots in figures 4.6 through 4.9. These plots depict a pattern whereby the estimated $\hat{v}_{it} + \hat{u}_i$ are larger than the $\hat{\gamma}_{it}$ on a more or less consistent basis. In fact, even when the time-varying inefficiency term v_{it} in the parametric model is estimated to be zero for all i and t, which is what occurred under the gross output specification with $v_{it} \sim \mathcal{N}^{+}(0, \sigma_v)$, the null hypothesis of FOSD is still not rejected in 3 of the 4 industries (the fabricated metal industry being the sole exception). Hence the difference in magnitude of the inefficiency estimates that is clearly

which is insufficient for a kernel regression with multiple predictors.

observed in figures 4.6 through 4.9 cannot be explained by the replacement of v_{it} by a time-varying frontier function $g_t(\mathbf{x})$ under the nonparametric framework. Rather, it is more likely that the greater flexibility of the kernel-based random effects estimator yields an improved fit of the frontier function which, in turn, enhances the accuracy of both the persistent and the time-varying components of firm-level inefficiency. Monte Carlo evidence to support this claim will be provided later on in Section 4.6.1.

4.5 Testing the plausibility of distributional assumptions

An important question that has not yet been answered in this chapter is the appropriateness of the assumption that the time-varying inefficiency and stochastic noise terms are distributed as either normal-half-normal or normal-exponential. Recall from Section 4.2.1 that this assumption is required for the separate identification of v_{it} and e_{it} under the parametric modelling framework; however, in the event that the distribution of $e_{it}-v_{it}$ has been erroneously specified, the ensuing estimates of firm-level inefficiency will be biased. It might therefore be prudent to scrutinize the distributional assumptions that underlie the parametric frontier model by means of one of the specification tests that is described in Amsler, Schmidt, and Wang (2011). Sections 4.5.1 and 4.5.2 summarize the two different approaches that these authors propose, namely the Kolmogorov-Smirnov and the Pearson χ^2 goodness of fit tests, while Section 4.5.3 points to evidence that favours rejection of the normal-half-normal and normal-exponential distributional assumptions in the parametric frontier

model.

4.5.1 Kolmogorov-Smirnov test

Let $F(\cdot)$ and $F_n(\cdot)$ respectively denote the normal-half-normal (or normal-exponential) cumulative distribution function and the empirical distribution function for the estimates of $e_{it} - v_{it}$. The KS statistic is given by:

$$KS = \sup_z |F(z) - F_n(z)|. \tag{4.15}$$

Given that the normal-half-normal and normal-exponential CDFs do not have known closed-form expressions, Amsler et al. (2011) recommend using tabulated quantiles from a simulated dataset and pre-determined values of the unknown parameters.[9] The authors also suggest that a wild bootstrapped $1-\alpha$ critical value for a KS test can be computed as follows: let $b = 1, ..., B$ denote the the bootstrap draw, and for $b = 1$, let $e_{it}^{(b)} - v_{it}^{(b)}$ denote sample values of the stochastic noise and time-varying inefficiency terms from the known distribution $F(\cdot)$. Next, define $y_{it}^{(b)} = \hat{\alpha}_0 + \mathbf{x}'_{it}\hat{\boldsymbol{\beta}} - \hat{u}_i + e_{it}^{(b)} - v_{it}^{(b)}$ and re-estimate $\mathbb{E}\left(y_{it}^{(b)}|\mathbf{x}_{it}\right)$, u_i, and $e_{it} - v_{it}$ using Heshmati and Kumbhakar's (1995) procedure from Section 4.2.1. One can then compare the empirical distribution $F_n(\cdot)$ of the latter with the known distribution $F(\cdot)$ of the $e_{it}^{(b)} - v_{it}^{(b)}$, and compute a Kolmogorov-Smirnov statistic $KS^{(b)}$ as in (4.15). Repeating this procedure B times yields the empirical distribution of the $KS^{(b)}$, from which the $1-\alpha$ critical value can be derived and compared with the Kolmogorov-Smirnov statistic that is based on the original data.

[9] For example, in the normal-half-normal case, a separate cdf needs to be tabulated for every unique value of $\lambda = \frac{\sigma_v}{\sigma_u}$.

4.5.2 Pearson χ^2 test

Once again, let $F(\cdot)$ and $F_n(\cdot)$ respectively denote the hypothesized CDF (i.e. normal-half-normal or normal-exponential) and the empirical distribution function for the estimates of $e_{it} - v_{it}$. Define the interval $\mathcal{I}_z = [z_{min}, z_{max}]$, where z_{min} is the lesser of $\sup\{z : F(z) = 0\}$ and $\sup\{z : F_n(z) = 0\}$, and z_{max} is the greater of $\inf\{z : F(z) = 1\}$ and $\inf\{z : F_n(z) = 1\}$. Amsler et al. (2011) write that the Pearson χ^2 test involves splitting $\mathcal{I}_z$ into k subintervals $\mathcal{I}_1, ..., \mathcal{I}_k$ and computing for each $j = 1, ..., k$:

$$\begin{aligned} O_j &= \sum_{i,t} 1(\hat{e}_{it} - \hat{v}_{it} \in \mathcal{I}_j) = \sum_{i,t} 1(z_{min_j} \leq \hat{e}_{it} - \hat{v}_{it} < z_{max_j}) \\ E_j &= NT \cdot [F(z_{max_j}) - F(z_{min_j})], \end{aligned} \tag{4.16}$$

where in (4.16), the endpoints of subinterval $\mathcal{I}_j$ are denoted by z_{min_j} and z_{max_j}.[10] Thus, O_j is the number of $\hat{e}_{it} - \hat{v}_{it}$ that are observed in the subinterval $\mathcal{I}_j$, while E_j is the number of $\hat{e}_{it} - \hat{v}_{it}$ that one would expect to observe in $\mathcal{I}_j$ if the null hypothesis of a normal-half-normal or normal-exponential distribution is true. The two expressions in (4.16) give rise to the following test statistic:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}. \tag{4.17}$$

Amsler et al. (2011) point out that if the density functions in (4.4) and (4.5) corresponding to the hypothesized CDF $F(\cdot)$ did not include any unknown parameters, the statistic in (4.17) would have a chi-squared distribution with

[10] The authors consider $k = 3$, $k = 5$, and $k = 10$ for sample sizes that range from $n = 50$ to $n = 500$, while the present chapter considers $k = 5$, $k = 10$, and $k = 25$ for sample sizes that range from $n = 1144$ to $n = 4862$. It turns out that the results are not very sensitive to the choice of k.

$k-1$ degrees of freedom. This would allow for a straightforward test of the null hypothesis that $F(\cdot)$ does not suffer from misspecification. However, inference is complicated by the fact that maximum likelihood estimates of the unknown parameters in (4.4) and (4.5) must be substituted into $F(\cdot)$ in order to compute (4.17). Instead, one can use a wild bootstrap procedure that is analogous to the one described in the previous subsection for the Kolmogorov-Smirnov goodness of fit test.

4.5.3 Results of specification tests

Tables 4.6 through 4.9 provide reason to doubt the validity of both the normal-half-normal and the normal-exponential distributional assumptions that underlie $e_{it} - v_{it}$ in the parametric frontier model. The p-values that are included in parentheses are nearly all equal to 0 up to several decimal places, which implies that the null hypothesis of a correctly specified parametric model is rejected at all conventional significance levels. The one exception that is worth noting is the value-added model for the furniture and finished wood products industry, where it cannot be rejected that the stochastic noise and time-varying inefficiency terms are distributed as normal-exponential. Altogether, both the Kolmogorov-Smirnov and the Pearson χ^2 tests yield similar conclusions, namely that in the context of the Colombian manufacturing data, the inclusion of v_{it} in the frontier model is potentially problematic. This serves as part of the justification for the fully nonparametric framework where intertemporal variation in firm-level inefficiency is conceptualized in terms of shifts in the frontier function; the advantage of this approach is that the intertemporal shifts can be identified separately from the random error e_{it} without making

any distributional assumptions. Of course, one must not lose sight of the fact that in the parametric model, v_{it} tends to be much smaller in magnitude than the persistent measure of inefficiency u_i whose distribution does not need to be known in advance. Thus, in practice, while the results of the specification tests are indeed significant, they pertain to only a relatively small fraction of firms' measured distance from the production frontier.

4.6 A semiparametric alternative to the parametric model

In the event that the specification tests described in Section 4.5 do not result in a rejection of the null hypothesis that $e_{it} - v_{it}$ is drawn from either a normal-half-normal or a normal-exponential distribution, it doesn't immediately follow that the parametric frontier model of Heshmati and Kumbhakar (1995) is the best available option. For example, it might be preferable to retain the parametric assumptions about the distribution of v_{it} and e_{it}, but to adopt a more flexible specification of the frontier function. In particular, consider the semiparametric framework below that constitutes somewhat of a "middle ground" between the approaches that are delineated in Sections 4.2.1 and 4.2.2:

$$y_{it} = g\left(\mathbf{x}_{it}\right) + e_{it} - v_{it} - u_i, \tag{4.18}$$

where the functional form of $g\left(\cdot\right)$ is unknown, the time subscript in (4.8) has been dropped in favour of a varying inefficiency term v_{it} with a known distribution, and both e_{it} and u_i are defined exactly as in Section 4.2.1. Estimation of (4.18) is very straightforward; it can be carried out by following the same procedure that is outlined in Section 4.2.1, with the exception that the para-

metric random effects regression is replaced by a kernel-based method. This ought to give rise to improved estimates of the frontier, factor elasticities, and firm-level inefficiency, even when the parametric density function that underlies the latter is correcly specified. In what follows, we illustrate this point by means of a simple Monte Carlo experiment.

4.6.1 Monte Carlo simulation

This section uses simulated data to evaluate the relative performance of the parametric and semiparametric frontier estimators when the distribution of the stochastic noise and varying inefficiency terms $e_{it} - v_{it}$ is correctly specified. The Monte Carlo experiment comprises $M = 500$ draws with a sample size of either $n = 500$ ($N = 50,\ T = 10$) or $n = 1000$ ($N = 100,\ T = 10$). Three different specifications of the frontier function are considered:

1. $y_{it} = 1 - x_{it} + e_{it} - v_{it} - u_i$
2. $y_{it} = \ln(1 + x_{it}) + e_{it} - v_{it} - u_i$
3. $y_{it} = 8(x_{it} - 0.5)^3 + e_{it} - v_{it} - u_i.$

Hence, the linear parametric model is correctly specified under the first scenario but it is misspecified under the other two. In each instance, x_{it} is i.i.d. $\mathcal{U}[0,1]$, while u_i, v_{it}, and e_{it} are respectively i.i.d. $\mathcal{U}[0,u_{max}]$, $\mathcal{N}^+(0,\sigma_v)$, and $\mathcal{N}(0,\sigma_e)$. The parameters u_{max} (and by extension, σ_u), σ_v, and σ_e are chosen so that the inefficiency and error terms collectively exhibit about a quarter of the variation of the frontier function. We assume that the persistent component of inefficiency varies twice as much as the time-varying component and four times as much as the stochastic noise term by setting $\sigma_u = 0.2\sigma_{f(x)}$,

$\sigma_v = 0.1\sigma_{f(x)}$, and $\sigma_e = 0.05\sigma_{f(x)}$, where $\sigma_{f(x)}$ denotes the standard deviation of the production frontier. The parametric and semiparametric approaches are evaluated based on the mean squared error of their respective estimates of i) the production frontier, ii) the factor elasticities $\frac{\partial y}{\partial x}(x_{it})$, and iii) firm-level inefficiency.

4.6.2 Finite-sample performance

We begin with the baseline case in which the parametric model in (4.1) does not suffer from any misspecification issues. A standard result from the statistics and econometrics literature is that a correctly-specified parametric model tends to outperform the semiparametric or nonparametric alternative, and this is precisely what is observed in the first stage of the Monte Carlo experiment. Table 4.10 provides a comparison of the median root mean squared error of the parametric and semiparametric frontier, elasticity, and inefficiency estimates. Under the baseline linear specification of the model, the parametric estimator yields the best fit overall. Nonetheless, the competing approach that makes use of the kernel-based random effects regression framework performs quite well, particularly vis-a-vis the measurement of firm-level inefficiency. In this case, there is zero difference in median RMSE between the two methods both when n=500 and n=1000. Meanwhile, when the production frontier is characterized by a logarithmic function, the parametric model is misspecified and the semiparametric estimator yields the best fit by a wide margin. For instance, in the middle portion of table 4.10, when the sample size n=1000, we observe that the median RMSE of the kernel-based frontier, elasticity, and inefficiency estimates are respectively 81, 88, and 42 percent lower than the parametric

alternatives. Even larger reductions of 95, 93, and 72 percent, respectively, are noted in the third simulated scenario when the frontier is cubic. An additional result that is omitted from table 4.10 but that is nevertheless worthy of note relates to the identification and estimation of the time-varying inefficiency term v_{it} when the frontier function is misspecified. It appears that under the third scenario where $f(x)$ is cubic, σ_v is frequently incorrectly estimated to be zero when the parametric frontier estimates are used in the log-likelihood function in (4.6). In particular, when n=500 and n=1000, $\hat{\sigma}_v = 0$ in 44 and 46 percent of the draws, respectively, while $\hat{\sigma}_v > 0$ in all 500 draws when the semiparametric approach is followed. This serves to partially explain why the estimates of firm-level inefficiency obtained under the parametric framework tend to be relatively imprecise.

4.7 Conclusion

This chapter has proposed a nonparametric estimation procedure that can be applied to stochastic production frontier models for panel data. It has shown that dispensing with parametric assumptions vis-a-vis the functional form of the frontier and the distribution of the convoluted error term has some key advantages, namely i) a greater capacity to account for intra-industry and intertemporal heterogeneity in factor elasticities and ii) more robust decomposition of firm-level inefficiency into its persistent and varying constituent parts. Using survey data from the Colombian manufacturing sector, this chapter has established that the parametric and nonparametric approaches give rise to substantially different estimates of firm-level inefficiency and in fact, these estimates are often characterized by a first-order stochastic dominance rela-

tionship. In addition, results from two different specification tests suggest that the parametric model is built on problematic distributional assumptions. Altogether, there is a lot to be gained from using kernel-based methods to evaluate the productive efficiency of firms in panel data settings.

References

Ackerberg, D., K. Caves, and G. Frazer (2006): "Structural identification of production functions," Mpra paper, University Library of Munich, Germany.

Aigner, D., C. A. K. Lovell, and P. Schmidt (1977): "Formulation and estimation of stochastic frontier production function models," *Journal of Econometrics*, 6, 21–37.

Amsler, C., P. Schmidt, and W. S. Wang (2011): "Goodness of fit tests in stochastic frontier models," *Journal of Productivity Analysis*, 35, 95–118.

Fan, Y., Q. Li, and A. Weersink (1996): "Semiparametric estimation of stochastic production frontier models," *Journal of Business and Economic Statistics*, 14, 460–468.

Fernandes, A. (2007): "Trade policy, trade volumes and plant-level productivity in Colombian manufacturing industries," *Journal of International Economics*, 71, 52–71.

Gandhi, A., S. Navarro, and D. Rivers (2016): "On the identification of production functions: How heterogeneous is productivity?" Manuscript.

Greene, W. (2005a): "Fixed and random effects in stochastic frontier models," *Journal of Productivity Analysis*, 23, 7–32.

——— (2005b): "Reconsidering heterogeneity in panel data estimators of the stochastic frontier model," *Journal of Econometrics*, 126, 269–303.

Hall, P. and L. Simar (2002): "Estimating a changepoint, boundary, or frontier in the presence of observation error," *Journal of the American Statistical Association*, 97, 523–534.

Hardaker, J. B., S. C. Kumbhakar, and G. Lien (2014): "Technical efficiency in competing panel data models: A study of Norwegian grain farming," *Journal of Productivity Analysis*, 41, 321–337.

Hayfield, T. and J. S. Racine (2008): "Nonparametric econometrics: The np package," *Journal of Statistical Software*, 27.

Henderson, D. J. and A. Ullah (2005): "A nonparametric random effects estimator," *Economics Letters*, 88, 403–407.

Heshmati, A. and S. C. Kumbhakar (1995): "Efficiency measurement in Swedish dairy farms: An application of rotating panel data," *American Journal of Agricultural Economics*, 77, 660–674.

Horrace, W. C. and C. F. Parmeter (2011): "Semiparametric deconvolution with unknown error variance," *Journal of Productivity Analysis*, 35, 129–141.

Jondrow, J., C. A. K. Lovell, I. S. Materov, and P. Schmidt (1982): "On the estimation of technical inefficiency in the stochastic frontier production function model," *Journal of Econometrics*, 19, 233–238.

Kealey, J. (2015): "Semiparametric estimation of a Cobb-Douglas production function with varying elasticity coefficients," Draft dissertation chapter, McMaster University.

Klecan, L., R. McFadden, and D. McFadden (1991): "A robust test for stochastic dominance," Department of economics working paper, MIT.

Kneip, A. and L. Simar (1996): "A general framework for frontier estimation with panel data," *Journal of Productivity Analysis*, 7, 187–212.

Kumbhakar, S. and C. Lovell (2000): *Stochastic Frontier Analysis*, Cambridge University Press.

Lee, J. and P. M. Robinson (2015): "Panel nonparametric regression with fixed effects," *Journal of Econometrics*, 188, 346–362.

Levinsohn, J. and A. Petrin (2003): "Estimating production functions using inputs to control for unobservables," *Review of Economic Studies*, 70, 317–341.

Li, Q., E. Maasoumi, and J. S. Racine (2009): "A nonparametric test for equality of distributions with mixed categorical and continuous data," *Journal of Econometrics*, 148, 186–200.

Li, Q. and J. S. Racine (2004): "Cross-validated local linear nonparametric regression," *Statistica Sinica*, 14, 485–512.

Linton, O., E. Maasoumi, and Y. J. Whang (2005): "Consistent testing for stochastic dominance under general sampling schemes," *Review of Economic Studies*, 72, 735–765.

Ma, S., J. Racine, and A. Ullah (2015): "Nonparametric regression-spline random effects models," Department of economics working papers, McMaster University.

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

Martins-Filho, C. and F. Yao (2009): "Nonparametric regression estimation with general parametric error covariance," *Journal of Multivariate Analysis*, 100, 309–333.

——— (2015): "Semiparametric stochastic frontier estimation via profile likelihood," *Econometric Reviews*, 34, 413–451.

Meeusen, W. and J. van den Broeck (1977): "Efficiency estimation fom Cobb-Douglas production functions with composed error," *International Economic Review*, 18, 435–444.

Meister, A. (2006): "Density estimation with normal measurement error with unknown variance," *Statistica Sinica*, 16, 195–211.

Olley, S. G. and A. Pakes (1996): "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 64, 1263–97.

Parmeter, C. and S. Kumbhakar (2014): "Efficiency analysis: A primer on recent advances," *Foundations and Trends in Econometrics*, 7, 191–385.

Parmeter, C. and J. S. Racine (2012): "Smooth constrained frontier analysis," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr.*, ed. by X. Chen and N. R. Swanson, New York: Springer-Verlag, 463–488.

Pombo, C. (1999): "Productividad industrial en Colombia: Una aplicacion

de numeros indices," *Revista de Economia de la Universidad del Rosario*, 107–139.

Roberts, M. and J. Tybout (1997): "The decision to export in Colombia: An empirical model of entry with sunk costs," *American Economic Review*, 87, 545–64.

Schmidt, P. and R. C. Sickles (1984): "Production frontiers and panel data," *Journal of Business and Economic Statistics*, 2, 367–374.

Su, L. and A. Ullah (2007): "More efficient estimation of nonparametric panel data models with random effects," *Economics Letters*, 96, 375–380.

Wooldridge, J. (2009): "On estimating firm-level production functions using proxy variables to control for unobservables," *Economics Letters*, 104, 112–114.

4.8 Appendix

4.8.1 Kernel regression with random effects

Martins-Filho and Yao (2009) propose a kernel-based nonparametric estimator for panel data models with random effects, such as the one that appears in (4.9) in Section 4.2.2:

$$y_{it} = h_t(\mathbf{x}_{it}) + e_{it} + \eta_i. \tag{4.19}$$

For a fixed $\mathbf{x} = (x_1, x_2, ..., x_q)'$, we define the $(2q+1) \times 1$ vector of predictors $\tilde{\mathbf{x}}_{it} = [1, x_{1it} - x_1, ..., x_{qit} - x_q, (x_{1it} - x_1)^2, ..., (x_{qit} - x_q)^2]'$, which is then used to construct the $NT \times (2q+1)$ matrix $\mathbf{X} = (\tilde{\mathbf{x}}_{11}, ..., \tilde{\mathbf{x}}_{1T}, ..., \tilde{\mathbf{x}}_{N1}, ..., \tilde{\mathbf{x}}_{NT})'$. We define the first-stage local quadratic estimator $\hat{\boldsymbol{\alpha}}^{1S}(\mathbf{x})$ as the solution to the following minimization problem:

$$\min_{\boldsymbol{\alpha}} (\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}(\mathbf{x}))' \mathbf{K}(\mathbf{x}) (\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}(\mathbf{x})), \tag{4.20}$$

where $\mathbf{K}(\mathbf{x})$ is a $NT \times NT$ kernel weighting matrix for a mixture of continuous and discrete data (see Li and Racine, 2004). The first element of $\hat{\boldsymbol{\alpha}}^{1S}(\mathbf{x})$ is the first-stage estimate of the conditional mean in (4.19), which can be subtracted from the output vector to obtain the residuals $\hat{\varepsilon}_{it} = y_{it} - \mathbf{e}_1' \hat{\boldsymbol{\alpha}}^{1S}(\mathbf{x}_{it})$, where $\mathbf{e}_1 = (1, 0, ..., 0)'$ is a $(2q+1) \times 1$ vector in which the first entry is a 1 and the remaining entries are 0's. We define $\hat{\sigma}_e^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\hat{\varepsilon}_{it} - \bar{\hat{\varepsilon}}_i\right)^2$ and $\hat{\sigma}_\eta^2 = \frac{1}{N} \sum_{i=1}^{N} \bar{\hat{\varepsilon}}_i^2 - \frac{\hat{\sigma}_\xi^2}{T}$, and the $NT \times NT$ matrix $\hat{\mathbf{P}}(\hat{\sigma}_e, \hat{\sigma}_\eta)$ that satisfies $\hat{\mathbf{P}}(\hat{\sigma}_e, \hat{\sigma}_\eta) \hat{\mathbf{P}}(\hat{\sigma}_e, \hat{\sigma}_\eta)' = \mathbb{I}_N \otimes (\hat{\sigma}_e \mathbb{I}_T + \hat{\sigma}_\eta \mathbf{1}_T \mathbf{1}_T')$, where $\mathbb{I}_N$, $\mathbb{I}_T$, and $\mathbf{1}_T$ respectively denote a $N \times N$ identity matrix, a $T \times T$ identity matrix, and a $T \times 1$ vector of 1's. Next, we construct the matrix $\hat{\mathbf{H}} = diag\left\{\hat{\rho}_{ii}^{-1}\right\}_{i=1}^{NT}$ using the diagonal elements $\hat{\rho}_{ii}$ of the matrix inverse $\hat{\mathbf{P}}^{-1}(\hat{\sigma}_e, \hat{\sigma}_\eta)$, and define the new variable

$\mathbf{z} = \hat{\mathbf{H}}\hat{\mathbf{P}}^{-1}\mathbf{y} + \left(\mathbb{I}_{NT} - \hat{\mathbf{H}}\hat{\mathbf{P}}^{-1}\right)\hat{\mathbf{h}}$, where $\hat{\mathbf{h}}$ is the vector of first-stage conditional mean estimates from (4.19). The second-stage conditional mean and gradient estimates for the random effects model $\hat{\alpha}(\mathbf{x}) = \left(\hat{h}_t(\mathbf{x}), \widehat{\nabla h}_t(\mathbf{x})', \widehat{\nabla^2 h}_t(\mathbf{x})'\right)'$ are given by the solution to the following:

$$\min_{\boldsymbol{\alpha}} \left(\mathbf{z} - \mathbf{X}\boldsymbol{\alpha}(\mathbf{x})\right)' \mathbf{K}(\mathbf{x}) \left(\mathbf{z} - \mathbf{X}\boldsymbol{\alpha}(\mathbf{x})\right), \tag{4.21}$$

where $\widehat{\nabla h}_t(\mathbf{x}) = \left(\widehat{\frac{\partial h_t}{\partial x_1}}(\mathbf{x}), ..., \widehat{\frac{\partial h_t}{\partial x_q}}(\mathbf{x})\right)'$ and $\widehat{\nabla^2 h}_t(\mathbf{x}) = \left(\widehat{\frac{\partial^2 h_t}{\partial x_1^2}}(\mathbf{x}), ..., \widehat{\frac{\partial^2 h_t}{\partial x_q^2}}(\mathbf{x})\right)'$.

4.8.2 Tables and figures

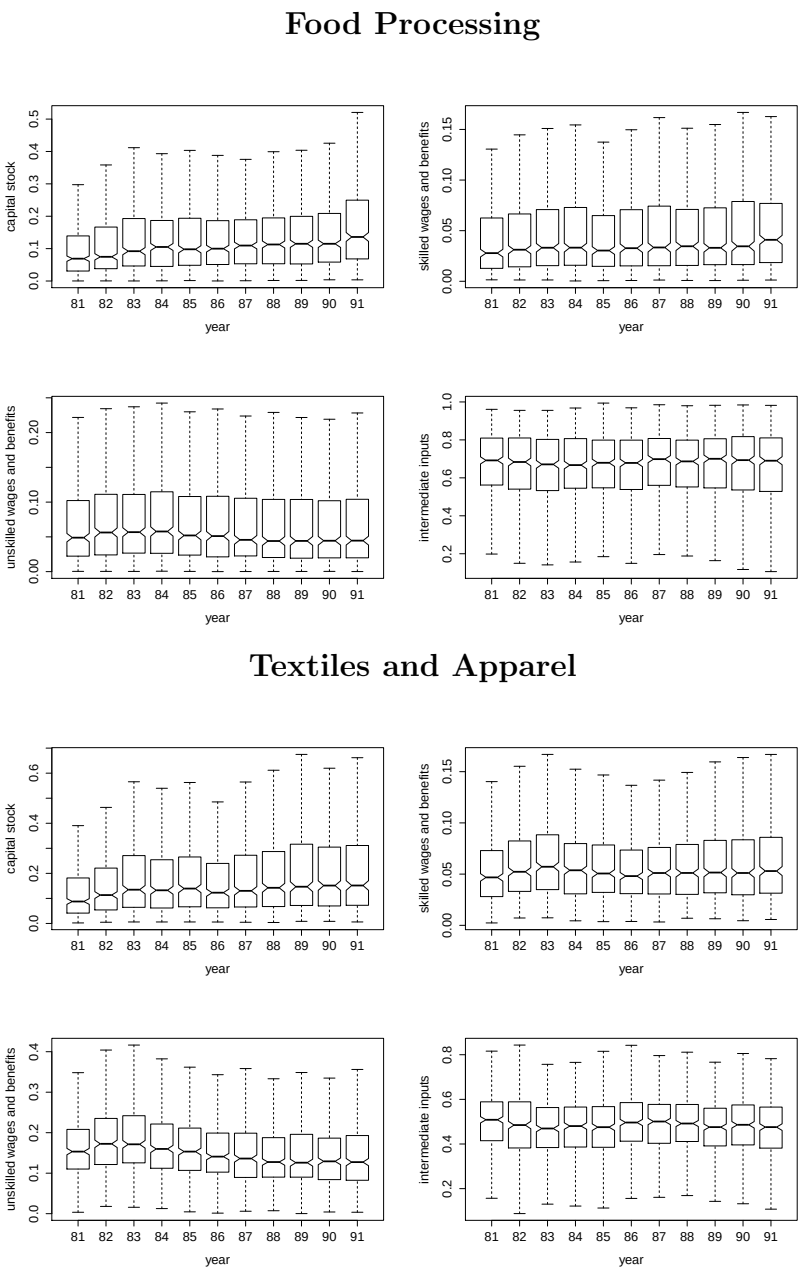

Figure 4.1: Boxplots of firm-level input expenditure shares of gross output over time

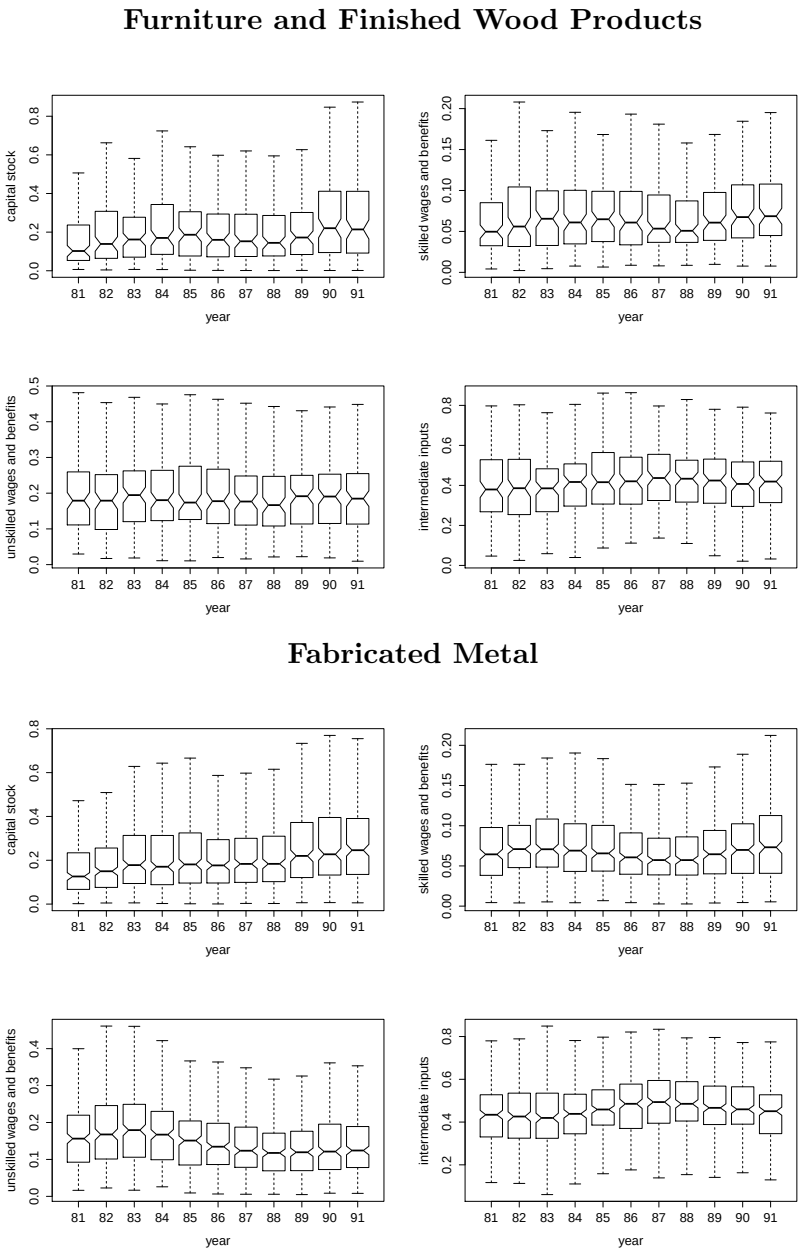

Figure 4.2: Boxplots of firm-level input expenditure shares of gross output over time

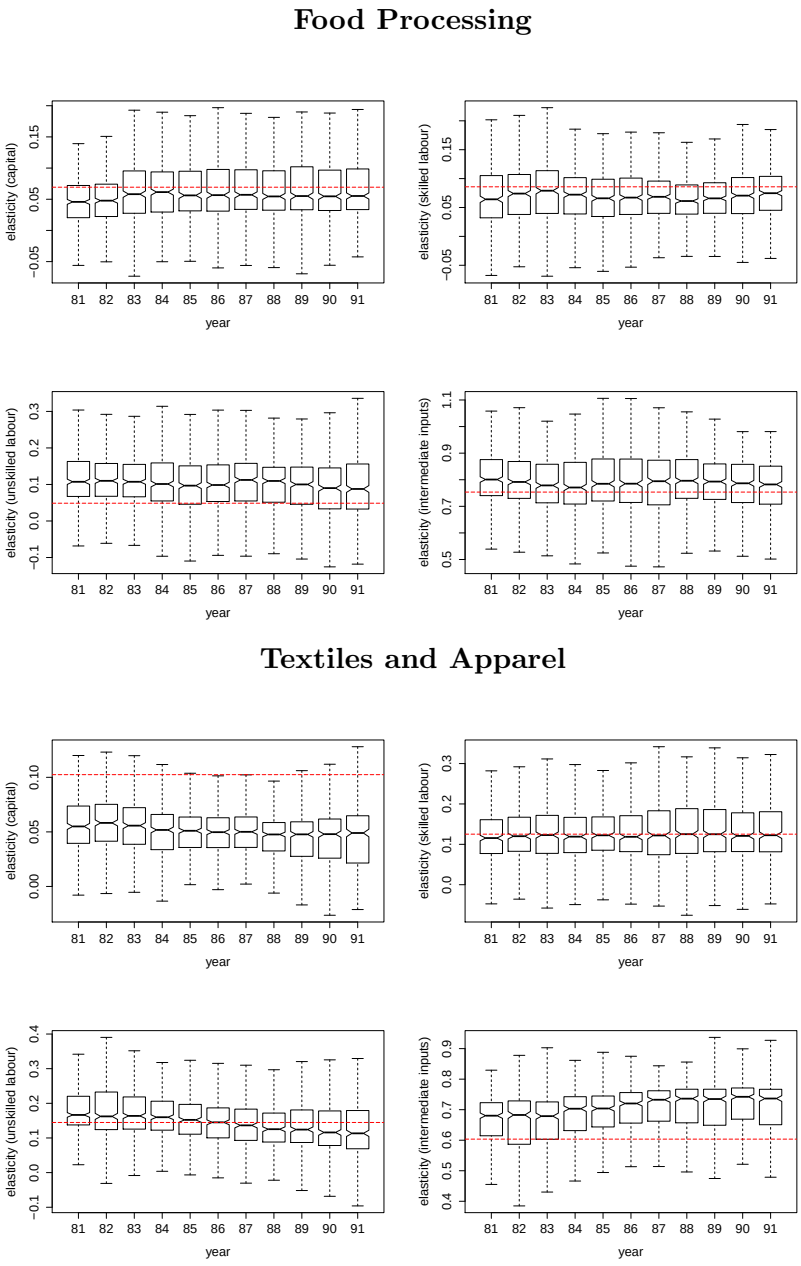

Figure 4.3: Elasticity boxplots for gross output model. Parametric estimates superimposed in red.

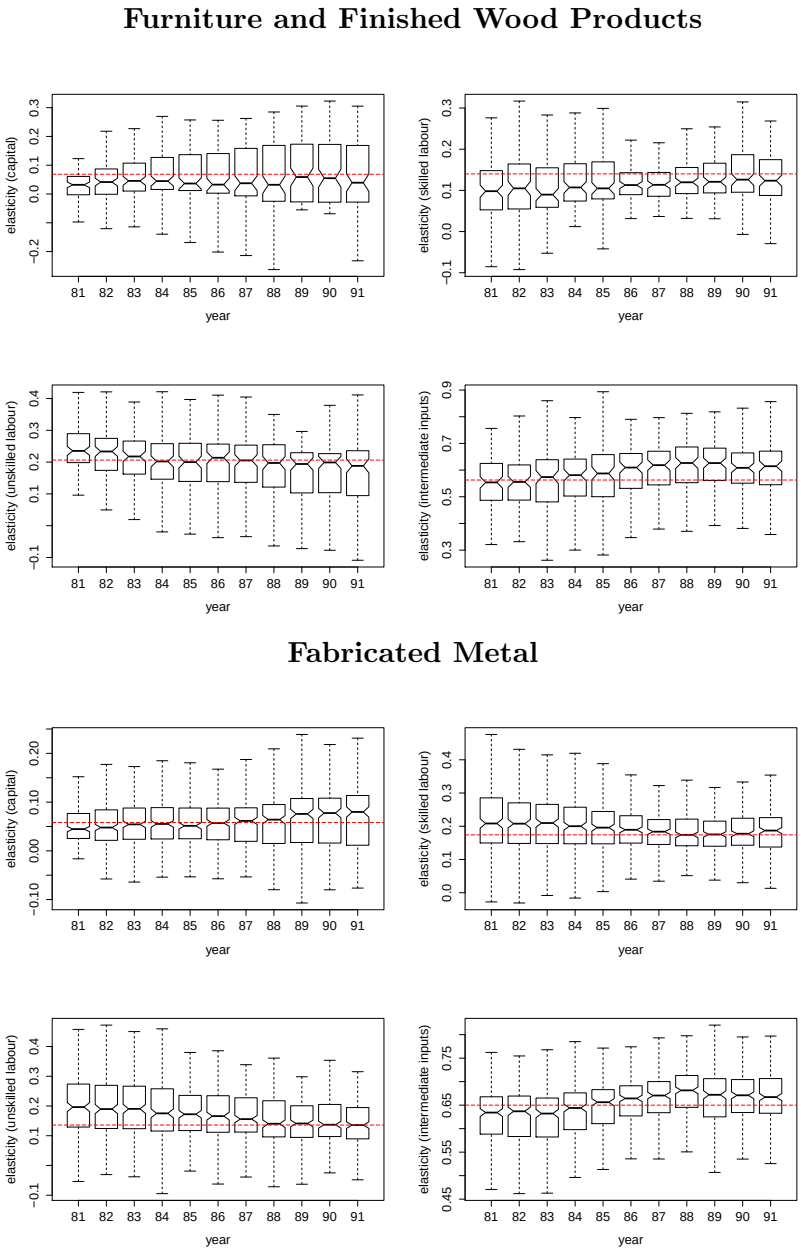

Figure 4.4: Elasticity boxplots for gross output model. Parametric estimates superimposed in red.

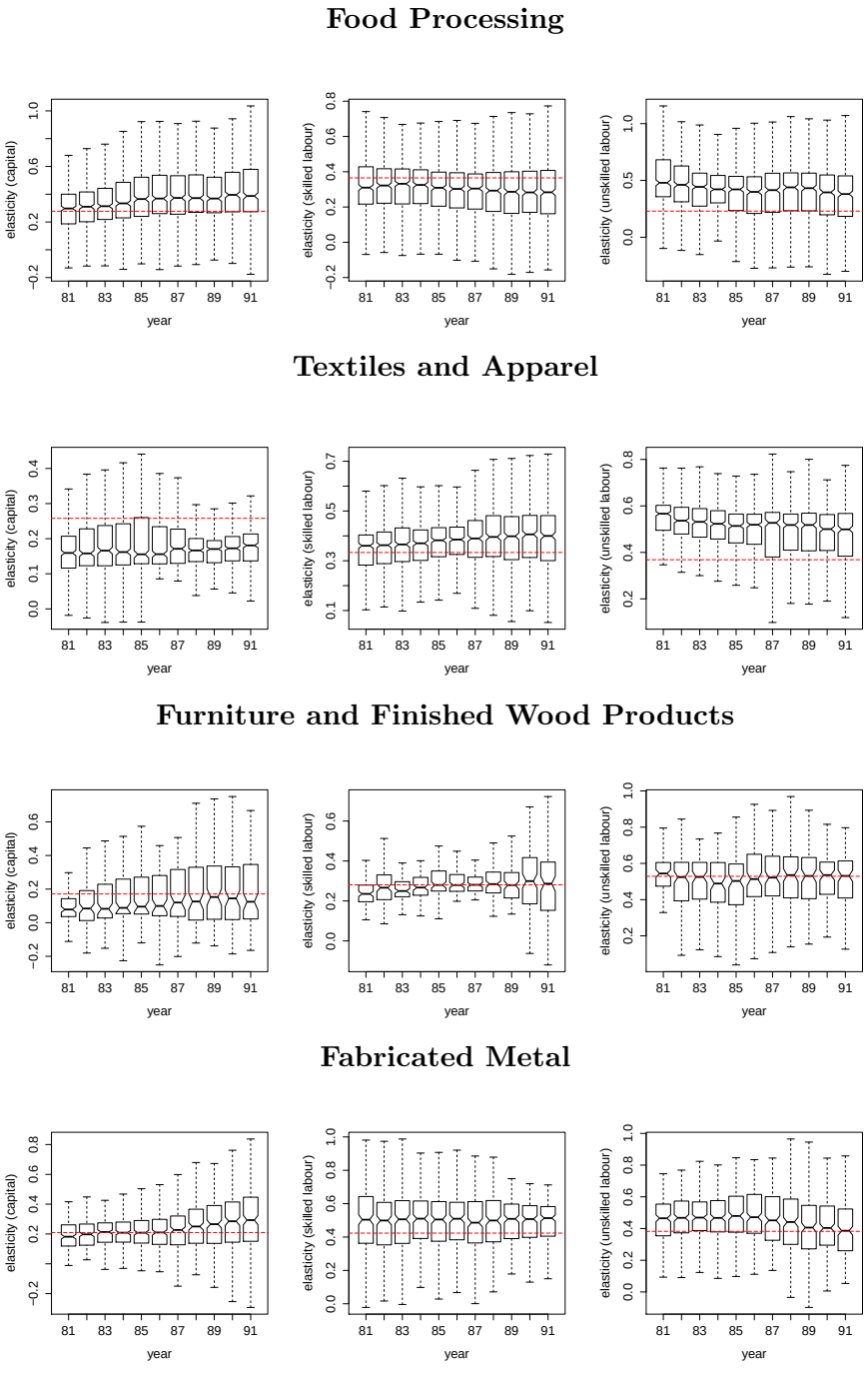

Figure 4.5: Elasticity boxplots for value-added model. Parametric estimates superimposed in red.

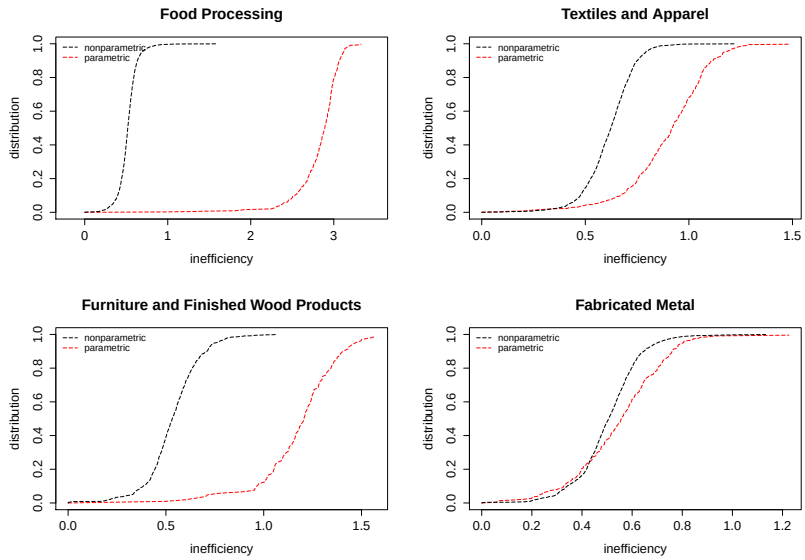

Figure 4.6: Distribution of inefficiency (gross output model with normal-half-normal assumption)

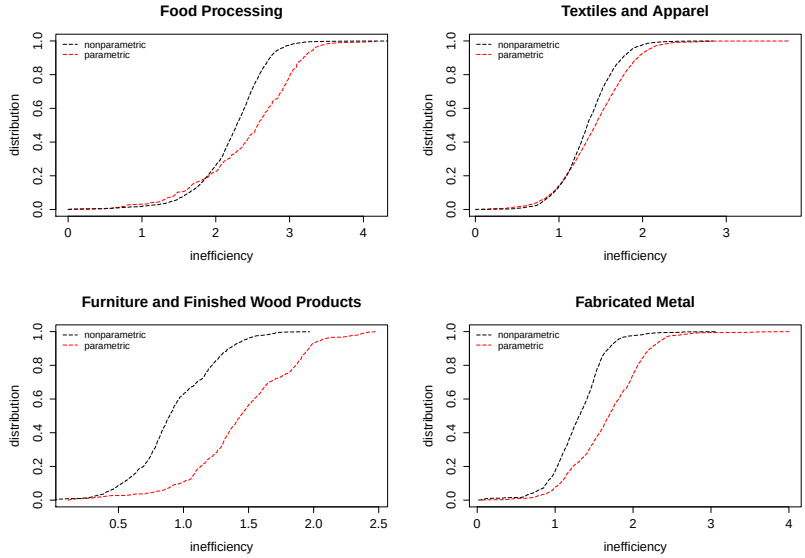

Figure 4.7: Distribution of inefficiency (value-added model with normal-half-normal assumption)

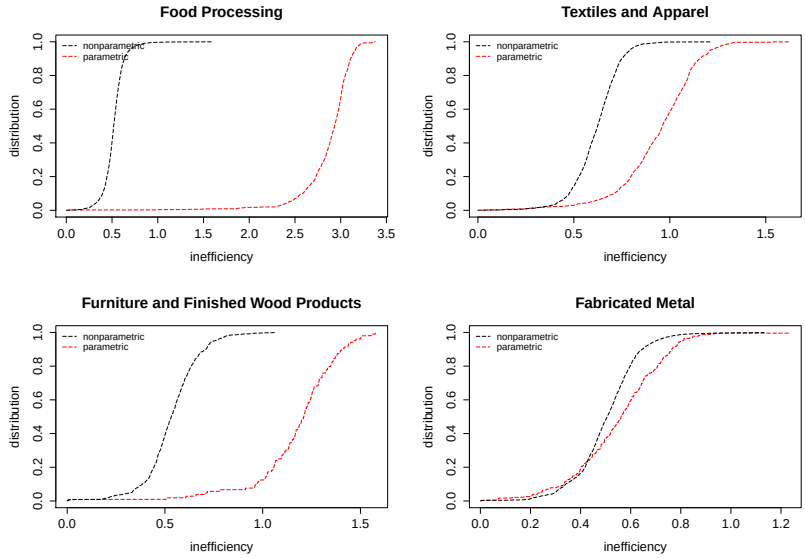

Figure 4.8: Distribution of inefficiency (gross output model with normal-exponential assumption)

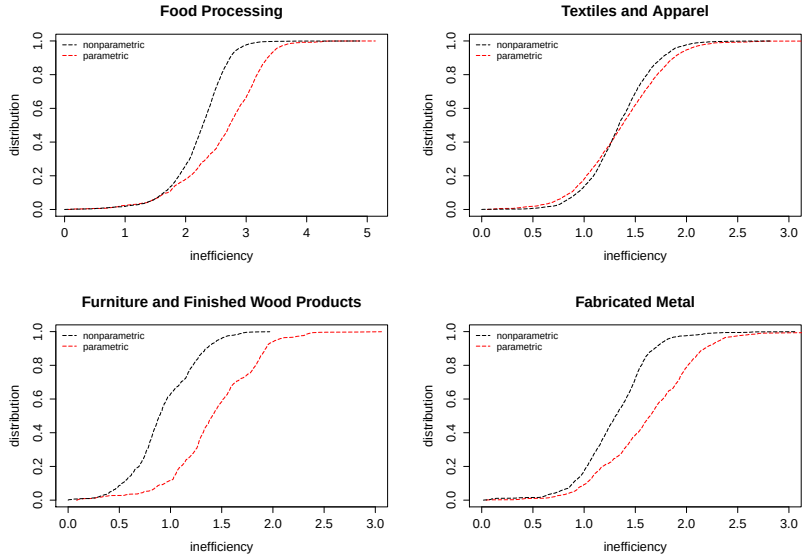

Figure 4.9: Distribution of inefficiency (value-added model with normal-exponential assumption)

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Capital				
Gross output	0.620	0.377	0.506	0.396
Value-added	0.210	0.519	0.597	0.426
Skilled labour				
Gross output	0.520	0.574	0.295	0.483
Value-added	-0.002	0.420	0.311	0.368
Unskilled labour				
Gross output	0.670	0.805	0.397	0.694
Value-added	0.244	0.350	0.309	0.438
Intermediates				
Gross output	0.740	0.628	0.671	0.538
Value-added	N/A	N/A	N/A	N/A

Table 4.1: Spearman correlations for nonparametric factor elasticities and corresponding expenditure shares

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Gross output	155.49	145.40	74.76	111.69
	(0.000)	(0.000)	(0.000)	(0.000)
Value-added	155.39	64.29	134.45	178.23
	(0.000)	(0.000)	(0.000)	(0.000)

Table 4.2: Li-Maasoumi-Racine (LMR) test statistics where the null hypothesis is equality of the parametric and nonparametric inefficiency estimates' densities. The parametric model assumes a half-normal distribution for v_{it}. Bootstrapped p-values are in parentheses.

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Gross output	1881.00	696.78	70.39	71.20
	(0.000)	(0.000)	(0.000)	(0.000)
Value-added	439.83	57.62	152.61	165.86
	(0.000)	(0.000)	(0.000)	(0.000)

Table 4.3: Li-Maasoumi-Racine (LMR) test statistics where the null hypothesis is equality of the parametric and nonparametric inefficiency estimates' densities. The parametric model assumes an exponential distribution for v_{it}. Bootstrapped p-values are in parentheses.

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Gross output	0.030	0.162	0.071	0.662
	(1.000)	(0.402)	(0.595)	(0.020)
Value-added	0.917	0.255	0.000	0.000
	(0.060)	(0.000)	(0.608)	(0.618)

Table 4.4: Linton-Maasoumi-Whang test statistics where the null hypothesis is that the parametric and nonparametric inefficiency estimates are characterized by a stochastic dominance relationship. The parametric model assumes a half-normal distribution for v_{it}. Bootstrapped p-values are in parentheses. B=199 bootstrap iterations are used for the food processing, textiles/apparel, and fabricated metal industries, while B=79 iterations are used for the furniture/wood products industry due to sample size limitations.

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Gross output	0.000	0.060	0.036	0.602
	(0.322)	(0.819)	(0.608)	(0.015)
Value-added	0.130	0.672	0.009	0.000
	(0.186)	(0.000)	(0.494)	(0.538)

Table 4.5: Linton-Maasoumi-Whang test statistics where the null hypothesis is that the parametric and nonparametric inefficiency estimates are characterized by a stochastic dominance relationship. The parametric model assumes an exponential distribution for v_{it}. Bootstrapped p-values are in parentheses. B=199 bootstrap iterations are used for the food processing, textiles/apparel, and fabricated metal industries, while B=79 iterations are used for the furniture/wood products industry due to sample size limitations.

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Gross output	0.130	0.080	0.062	0.084
	(0.000)	(0.000)	(0.003)	(0.000)
Value-added	0.088	0.050	0.067	0.071
	(0.000)	(0.000)	(0.000)	(0.000)

Table 4.6: Kolmogorov-Smirnov (KS) test statistics where the null hypothesis is a normal-half-normal distribution for $e_{it} - v_{it}$. Bootstrapped p-values are in parentheses.

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Gross output	0.132	0.084	0.066	0.074
	(0.000)	(0.000)	(0.000)	(0.000)
Value-added	0.088	0.048	0.034	0.073
	(0.000)	(0.000)	(0.313)	(0.000)

Table 4.7: Kolmogorov-Smirnov (KS) test statistics where the null hypothesis is a normal-exponential distribution for $e_{it} - v_{it}$. Bootstrapped p-values are in parentheses.

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Gross output				
$k = 5$	992.14	426.36	40.79	148.45
	(0.000)	(0.000)	(0.000)	(0.000)
$k = 10$	1016.24	434.78	67.15	184.14
	(0.000)	(0.000)	(0.000)	(0.000)
$k = 25$	1313.97	613.02	97.56	257.11
	(0.000)	(0.000)	(0.010)	(0.000)
Value-added				
$k = 5$	499.90	142.34	42.19	166.10
	(0.000)	(0.000)	(0.000)	(0.000)
$k = 10$	515.29	185.28	51.40	131.36
	(0.000)	(0.000)	(0.000)	(0.000)
$k = 25$	638.49	162.77	95.02	171.39
	(0.000)	(0.000)	(0.008)	(0.000)

Table 4.8: Pearson χ^2 test statistics where the null hypothesis is a normal-half-normal distribution for $e_{it} - v_{it}$. Bootstrapped p-values are in parentheses.

	Food Processing	Textiles/Apparel	Furniture/Wood Products	Fabricated Metal
Gross output				
$k = 5$	879.66	410.00	78.27	156.28
	(0.000)	(0.000)	(0.000)	(0.000)
$k = 10$	1074.16	431.49	62.24	205.22
	(0.000)	(0.000)	(0.003)	(0.000)
$k = 25$	1060.51	541.10	83.99	214.21
	(0.000)	(0.000)	(0.013)	(0.000)
Value-added				
$k = 5$	388.87	104.76	35.20	67.77
	(0.000)	(0.000)	(0.000)	(0.000)
$k = 10$	518.82	137.33	17.81	117.32
	(0.000)	(0.000)	(0.343)	(0.000)
$k = 25$	554.85	166.68	94.77	174.67
	(0.000)	(0.000)	(0.003)	(0.000)

Table 4.9: Pearson χ^2 test statistics where the null hypothesis is a normal-exponential distribution for $e_{it} - v_{it}$. Bootstrapped p-values are in parentheses.

		N=50, T=10		N=100, T=10	
		Parametric	Semiparametric	Parametric	Semiparametric
1.	$f(x) = 1 - x$				
	frontier	0.0066	0.0073	0.0043	0.0048
	elasticity	0.0035	0.0148	0.0026	0.0113
	inefficiency	0.0150	0.0150	0.0140	0.0140
2.	$f(x) = \ln(1 + x)$				
	frontier	0.0183	0.0052	0.0184	0.0035
	elasticity	0.1400	0.0230	0.1400	0.0170
	inefficiency	0.0160	0.0100	0.0164	0.0096
3.	$f(x) = 8(x - 0.5)^3$				
	frontier	0.1570	0.0110	0.1550	0.0070
	elasticity	1.9600	0.1700	1.9600	0.1400
	inefficiency	0.0690	0.0200	0.0640	0.0180

Table 4.10: Median root mean squared error of the parametric and semiparametric frontier, elasticity, and inefficiency estimates based on M=500 draws with sample sizes n=500 and n=1000.

Chapter 5

Trade policy reform and firm-level productivity growth: Does the choice of production function matter?

5.1 Introduction

Contemporary theories of international trade tend to advance the point of view that import competition is beneficial for the productivity of domestic firms. From this perspective, one of the key advantages of a liberalized trade policy environment is that, by expanding the availability of foreign-produced goods, it encourages innovation among local producers who do not wish to see their market share erode. This, in turn, has a modernizing effect on the home country's industrial landscape. The need for empirical validation of the aforementioned theoretical stance has, in recent decades, given rise to a vast literature that is concerned with estimating the influence that trade barriers have on the dynamics of firm productivity. Of course, it is important to recognize that any serious discussion pertaining to the productivity of firms needs to be grounded in a well-thought-out methodological framework that

allows for proper identification of an underlying production function. While many empirical researchers acknowledge this fact, they rarely give sufficient consideration to the sensitivity of their findings to their chosen strategy for identifying and estimating firm productivity. Hence this chapter considers three different identification strategies that are commonly employed for the estimation of production functions, namely those of Levinsohn and Petrin (2003; henceforth LP), Ackerberg, Caves, and Frazer (2006; henceforth ACF), and Gandhi, Navarro, and Rivers (2016; henceforth GNR), and examines whether they yield consistent conclusions vis-à-vis firm-level productivity growth during periods of trade liberalization. Using data from the Colombian manufacturing sector, which has appeared in a number of related studies in the past, we find that switching from a "control function" Cobb-Douglas specification to a more flexible nonparametric framework tends to alter our findings regarding certain industries' experience with trade policy reforms.

A fair amount of evidence can be found in the empirical literature of a negative association between barriers to trade and firm productivity. For instance, Tybout and Westbrook (1995), Pavcnik (2002), Schor (2004), Fernandes (2007), Topalova and Khandelwal (2011), and Hu and Liu (2014) demonstrate that the liberalization of trade policy has generally coincided with productivity growth at the firm-level in Mexico, Chile, Brazil, Colombia, India, and China, respectively. The empirical focus of these studies tends to be the conditional mean of firm productivity, given different levels of trade protection; that is, most authors employ linear regression methods to evaluate whether there exists a rather general relationship between trade policy on the one hand, and the conditional expectation of firm productivity on the other. However, results that are obtained using standard linear regression techniques fail to shed light on whether different types of firms, ranging from the least to the most efficient producers of a particular good, exhibit similar responses to changes in the policy environment. Thus, in the present chapter, we opt for a quantile re-

gression approach that is better able to reflect trends in the *distribution* of firm productivity, as opposed to its conditional mean, during Colombia's era of liberalization. From a theoretical point of view, it makes sense to focus on outcomes at different quantiles because there is likely some intra-industry variation in the effect that competition from trade has on innovation behaviour and productivity. Indeed, Melitz (2003) posits that open trade enhances productivity through three distinct channels, namely i) reallocation of resources and market share from inefficient to efficient producers, ii) market exit on the part of inefficient firms, and iii) market entry on the part of efficient firms.[1] Melitz and Polanec (2015) build on previous work by Olley and Pakes (1996) and propose a decomposition procedure that allows for the empirical isolation of these contributing factors to aggregate productivity growth. We apply this methodology to each of the LP, ACF, and GNR productivity estimates and set out to identify any overlap that exists in our results. It turns out that our decomposition of industry-level productivity growth into the effects of market share reallocation among incumbents, exit of inefficient producers, and entry of productive firms is quite sensitive to the chosen identification strategy for estimation of the production function.

The remainder of the chapter is structured as follows: Section 5.2 provides a comprehensive summary of the three different methods that we employ to estimate firm-level productivity in the Colombian manufacturing sector. Section 5.3 describes the input, output, and trade policy data that is used in the analysis in Section 5.4, where we discuss the coefficient estimates that we obtain under several specifications of our quantile regression model, and the results of the Melitz-Polanec decomposition exercise that we perform for a long list of manufacturing industries. Section 5.5

[1] Thus, the testable implications of Melitz's (2003) theoretical framework relate primarily to the distribution of productivity within a particular industry insofar as entry, exit, and market share reallocation among incumbents will tend to have a pronounced effect on productivity growth near the left and right tails of the distribution of firms.

concludes.

5.2 A review of methods for estimating firm productivity

In this section, we provide a detailed overview of three different strategies for the identification and estimation of firm-level productivity. These approaches, which are presented in the chronological order of their appearance in the productivity literature, were originally proposed by Levinsohn and Petrin (2003), Ackerberg et al. (2006), and Gandhi et al. (2016), and are now in widespread use in a number of different subfields of empirical economics. In what follows, we adopt the convention whereby lower-case (upper-case) letters are used to denote the log (level) values of the variables in the production model.

5.2.1 Levinsohn and Petrin's control function method

Consider a logarithmically-transformed Cobb-Douglas production function:

$$y_{it} = \alpha_k k_{it} + \alpha_l l_{it} + \alpha_m m_{it} + \omega_{it} + \varepsilon_{it}, \tag{5.1}$$

where y_{it} is the log of firm i's gross output in period t, k_{it} is the log of the capital stock, l_{it} is the log of quantity of labour employed by the firm, and m_{it} is the log of an intermediate input variable comprising raw materials and energy consumption. Firm-level productivity is denoted by ω_{it} and ε_{it} is a random error term. Levinsohn and Petrin (2003) propose a "control function" approach whereby the firm's intermediate input demand is a function of its capital stock and its level of productivity:

$$m_{it} = m\left(k_{it}, \omega_{it}\right). \tag{5.2}$$

Assuming that the function $m(\cdot)$ is strictly increasing in ω_{it} holding k_{it} fixed, one can invert (5.2) to obtain an expression for firm-level productivity:

$$\omega_{it} = m^{-1}(k_{it}, m_{it}). \tag{5.3}$$

Inserting (5.3) into (5.1) yields:

$$\begin{aligned} y_{it} &= \alpha_k k_{it} + \alpha_l l_{it} + \alpha_m m_{it} + m^{-1}(k_{it}, m_{it}) + \varepsilon_{it} \\ &= \alpha_l l_{it} + \theta(k_{it}, m_{it}) + \varepsilon_{it}, \end{aligned} \tag{5.4}$$

where $\theta(k_{it}, m_{it}) = \alpha_k k_{it} + \alpha_m m_{it} + m^{-1}(k_{it}, m_{it})$. One can specify $\theta(k_{it}, m_{it})$ as a third-order polynomial[2] in k_{it} and m_{it} and estimate (5.4) by means of an ordinary least squares regression. This yields an estimate of the elasticity of output with respect to labour, $\hat{\alpha}_l$.

Next, Levinsohn and Petrin's (2003) framework assumes that firm-level productivity evolves according to a first-order Markov process:

$$\omega_{it} = g(\omega_{it-1}) + \eta_{it}, \tag{5.5}$$

where η_{it} can be interpreted as an unanticipated productivity shock. Using the fitted values $\hat{\theta}(k_{it}, m_{it})$ from the regression in (5.4), one can obtain the following expression for ω_{it}:

$$\omega_{it}(\alpha_k, \alpha_m) = \hat{\theta}(k_{it}, m_{it}) - \alpha_k k_{it} - \alpha_m m_{it}. \tag{5.6}$$

Lagged productivity, $\omega_{it-1}(\alpha_k, \alpha_m)$, is analogously defined. We specify (5.5) as a third-order polynomial without an intercept $\omega_{it} = \rho_1 \omega_{it-1} + \rho_2 \omega_{it-1}^2 + \rho_3 \omega_{it-1}^3 + \eta_{it}$

[2]We make the same functional form assumptions as LP, who propose a third-order polynomial specification of $\theta(k_{it}, m_{it})$.

and estimate ρ_1, ρ_2, and ρ_3 for given values of α_k and α_m, which allows us to write the unanticipated productivity shock as a function of the unknown elasticity parameters $\eta_{it}(\alpha_k, \alpha_m)$. Levinsohn and Petrin (2003) use the following moment condition to identify the elasticity of output with respect to capital and intermediate inputs:

$$\mathbb{E}\left[\eta_{it}(\alpha_k, \alpha_m) | k_{it}, m_{it-1}\right] = 0. \tag{5.7}$$

Finally, once $\hat{\alpha}_k$ and $\hat{\alpha}_m$ have been obtained using the sample analogue of the moment condition in (5.7), they can be plugged into (5.6) to obtain firm i's period-t productivity, $\hat{\omega}_{it}$.

5.2.2 Ackerberg, Caves, and Frazer's value-added model

Ackerberg et al. (2006) point out that Levinsohn and Petrin's (2003) approach suffers from a multicollinearity issue stemming from the likelihood that a firm's labour and intermediate input decisions are both influenced by its level of productivity. They show how this can complicate estimation of α_l in the partially linear model that is depicted in (5.4), and as an alternative, they propose the following value-added Cobb-Douglas production model:

$$va_{it} = \alpha_k k_{it} + \alpha_l l_{it} + \omega_{it} + \varepsilon_{it}, \tag{5.8}$$

where now, va_{it} denotes the log of firm i's value-added output in period-t. The right-hand side of (5.8) is the same as in (5.1), with the exception that the intermediate input variable m_{it} has been omitted. Ackerberg et al. (2006) use the same control

function as Levinsohn and Petrin (2003) that appears in (5.3), and rewrite (5.8) as:

$$\begin{aligned} va_{it} &= \alpha_k k_{it} + \alpha_l l_{it} + m^{-1}\left(k_{it}, m_{it}\right) + \varepsilon_{it} \\ &= \phi\left(k_{it}, l_{it}, m_{it}\right) + \varepsilon_{it}. \end{aligned} \tag{5.9}$$

Note that the central difference between the current approach and the one described in Section 5.2.1 lies in the specification of $\phi\left(k_{it}, l_{it}, m_{it}\right)$ in (5.9) as opposed to that of $\theta\left(k_{it}, l_{it}, m_{it}\right)$ in (5.4). Once again, $\phi\left(k_{it}, l_{it}, m_{it}\right)$ can be specified as a third-order polynomial in k_{it}, l_{it}, and m_{it} and estimated via OLS.[3] Productivity can then be written as $\omega_{it}\left(\alpha_k, \alpha_l\right) = \hat{\phi}\left(k_{it}, l_{it}, m_{it}\right) - \alpha_k k_{it} - \alpha_l l_{it}$ and the productivity shock η_{it} in (5.5) can be expressed in terms of the unknown elasticity parameters $\eta_{it}\left(\alpha_k, \alpha_l\right)$ by following the same procedure that was described in the previous subsection. Finally, Ackerberg et al. (2006) use the following moment condition to identify α_k and α_l:

$$\mathbb{E}\left[\eta_{it}\left(\alpha_k, \alpha_l\right) | k_{it}, l_{it-1}\right] = 0. \tag{5.10}$$

Again, once the elasticity parameters $\hat{\alpha}_k$ and $\hat{\alpha}_l$ have been obtained using the sample analogue of the moment condition in (5.10), firm-level productivity is computed as $\hat{\omega}_{it} = \hat{\phi}\left(k_{it}, l_{it}, m_{it}\right) - \hat{\alpha}_k k_{it} - \hat{\alpha}_l l_{it}$.

5.2.3 Gandhi, Navarro, and Rivers' nonparametric identification strategy

Gandhi et al. (2016) show how one can estimate a production function whose un-

[3] ACF estimate $\phi\left(k_{it}, l_{it}, m_{it}\right)$ by means of a kernel regression, but we follow the approach of LP who propose a third-order polynomial specification for the first-stage regression. The correlation coefficient of the fitted $\hat{\phi}\left(k_{it}, l_{it}, m_{it}\right)$ obtained under the two different approaches lies between 0.980 and 0.996 in all industries in the sample.

derlying functional form is unknown:

$$Y_{it} = F\left(K_{it}, L_{it}, M_{it}\right) e^{\omega_{it}+\varepsilon_{it}}, \tag{5.11}$$

where the upper-case Y_{it}, K_{it}, L_{it}, and M_{it} denote the output, capital stock, labour, and intermediate input variables in level form. Meanwhile, the productivity and error terms are once again denoted by ω_{it} and ε_{it}, respectively. Gandhi et al.'s (2016) approach makes use of the firm's first-order condition for its choice of intermediate inputs:

$$p_M = p_Y F_M\left(K_{it}, L_{it}, M_{it}\right) e^{\omega_{it}} E\left[e^{\varepsilon_{it}}\right], \tag{5.12}$$

where p_M and p_Y are respectively the intermediate input and final output prices and $F_M\left(K_{it}, L_{it}, M_{it}\right)$ is the partial derivative of the production function with respect to the intermediate input variable. Next, it can be shown that if one subtracts the log of (5.11) from the log of (5.12) and subsequently adds the log of M_{it} to both sides of the resulting expression, one obtains:

$$\ln\left(\frac{p_M M_{it}}{p_Y Y_{it}}\right) = \ln\left(\frac{F_M\left(K_{it}, L_{it}, M_{it}\right) M_{it}}{F\left(K_{it}, L_{it}, M_{it}\right)} E\left(e^{\varepsilon_{it}}\right)\right) - \varepsilon_{it}, \tag{5.13}$$

where (5.13) reflects the well-known theoretical relationship that ought to exist between intermediate input expenditures' share of total revenue and the elasticity of output with respect to intermediate inputs. The left-hand side of (5.13) can be computed using firm-level input expenditure and revenue data, while the expression in parentheses on the right-hand side can be approximated by a second-order polynomial[4] in k_{it}, l_{it}, and m_{it} (lower case letters denote the logs of the input variables). The equation can then be estimated by means of a non-linear least squares regression, and this yields estimates of ε_{it}, $E\left(e^{\varepsilon_{it}}\right)$, and $\frac{F_M(K_{it},L_{it},M_{it})M_{it}}{F(K_{it},L_{ji},M_{it})}$.

[4]This is the same specification that is used by GNR.

As a next step in the process of identifying a firm's production function, Gandhi et al. (2016) make use of the equality $\frac{F_M(K_{it},L_{it},M_{it})}{F(K_{it},L_{it},M_{it})} = \frac{\partial}{\partial M_{it}} \ln F(K_{it}, L_{it}, M_{it})$. Integrating both sides of this expression gives us

$$\int \frac{F_M(K_{it}, L_{it}, M_{it}) M_{it}}{F(K_{it}, L_{it}, M_{it})} \frac{\mathrm{d}M_{it}}{M_{it}} = \ln F(K_{it}, L_{it}, M_{it}) + \mathscr{C}(K_{it}, L_{it}). \tag{5.14}$$

Given that $\frac{F_M(K_{it},L_{it},M_{it})M_{it}}{F(K_{it},L_{it},M_{it})}$ has already been identified and estimated in (5.13), the expression above makes it possible to identify $\ln F(K_{it}, L_{it}, M_{it})$ up to a constant of integration, which Gandhi et al. (2016) denote by $\mathscr{C}(K_{it}, L_{it})$.[5] Combining (5.14) and the log of (5.11), the firm-level productivity term ω_{it} satisfies the following equality:

$$\omega_{it} = \ln Y_{it} - \int \frac{F_M(K_{it}, L_{it}, M_{it}) M_{it}}{F(K_{it}, L_{it}, M_{it})} \frac{\mathrm{d}M_{it}}{M_{it}} - \varepsilon_{it} + \mathscr{C}(K_{it}, L_{it}). \tag{5.15}$$

For the sake of notational simplicity, the above expression is rewritten as

$$\omega_{it} = \mathcal{Y}_{it} + \mathscr{C}(K_{it}, L_{it}), \tag{5.16}$$

where $\mathcal{Y}_{it}$ is shorthand for the more cumbersome $\ln Y_{it} - \int \frac{F_M(K_{it},L_{it},M_{it})M_{it}}{F(K_{it},L_{it},M_{it})} \frac{\mathrm{d}M_{it}}{M_{it}} - \varepsilon_{it}$. Lagged productivity, ω_{it-1}, is analogously defined. The constant of integration is modelled as a second-order polynomial in k_{it} and l_{it}.[6] Once again, we can follow the same procedure that was described in Sections 5.2.1 and 5.2.2 and model the evolution of ω_{it} as a first-order Markov process $\omega_{it} = \rho_1\omega_{it-1} + \rho_2\omega_{it-1}^2 + \rho_3\omega_{it-1}^3 + \eta_{it}$. The moment condition $E(\eta_{it}|K_{it}, L_{it}, \mathcal{Y}_{it-1}, K_{it-1}, L_{it-1}) = 0$ identifies the parameters in $\mathscr{C}(K_{it}, L_{it})$, yielding an estimate of firm-level productivity ω_{it}.

[5] Note that the integral has a straightforward closed-form solution because a second-order polynomial approximation was used to estimate $\frac{F_M(K_{it},L_{it},M_{it})M_{it}}{F(K_{it},L_{it},M_{it})}$.

[6] Again, this is the specification that is used by GNR.

5.3 Data

The dataset that underlies the analysis in Section 5.4 is taken from a census of Colombian manufacturers whose participants include all plants with 10 or more employees over the 11-year period 1981-1991. It consists of more than 61,000 observations on nearly 11,000 plants in 22 different industries. Note that while industries are classified according to their 3-digit ISIC code, they can be further subdivided on the basis of the 71 unique 4-digit ISIC codes that appear in the sample. The primary advantage of using the Colombian manufacturing data is that it has appeared in previous empirical studies that examine the relationship between trade and firm-level productivity (Roberts and Tybout, 1997; Fernandes, 2007). The gross ouput, value-added, capital stock, and intermediate input variables are all expressed in thousands of Colombian pesos, and are deflated using an industry-by-year price index that is found in the data.[7] Intermediate inputs, which are included in the production functions of Levinsohn and Petrin (2003) and Gandhi et al. (2016) but absent from that of Ackerberg et al. (2006), are defined as the total amount of energy and raw materials consumed by a plant in a given year. A plant's value-added production is therefore obtained by subtracting its intermediate input consumption from its gross output. Meanwhile, the labour variable is expressed as the total number of workers that are on a plant's payroll, but with the slight modification that unskilled and skilled labourers are weighted by the ratio of their respective median salaries.

The trade policy predictor that appears in our regression model is measured in two different ways. First, we use the Colombian government's import tariff schedule that is available for each of the 71 unique 4-digit ISIC codes that are represented in the census. For the 11-year period that runs from 1981 to 1981, tariff data is miss-

[7] In particular, both the nominal and the real value of production is recorded for each observation in the panel of manufacturing plants and hence, the ratio of these two variables serves as an industry-level price index.

ing for 1982 and 1989-1991, and so the first specification of the regression model is estimated using a 7-year subsample of the original dataset. In addition to the tariff data, we also use the effective rate of protection (ERP) as a trade policy indicator. This is intended to reflect the dual impact of protectionism, i.e. reduced competition from abroad on the one hand and increased imported input costs on the other. The ERP is computed as $\frac{va_d - va_w}{va_w}$, where va_d and va_w respectively denote manufacturers' value-added under distorted domestic (d) and undistorted world (w) prices. The effective rate of protection data is available for 22 unique 3-digit ISIC codes for the years 1981, 1984, 1985, 1990, and 1991, and so once again, the regressions that include the ERP as a predictor are only carried out on a 5-year subsample of the data. Tables 5.1 and 5.2 shed some light on the extent to which Colombia's trade policy regime underwent reform during the period that is under consideration. Minimum and maximum tariff and ERP values are reported for each of the 3-digit and 4-digit industries that are covered by the sample. In many instances, there is substantial liberalization, with some industries experiencing a 50 to 60 percentage point decrease in import tariffs between the mid-1980s and the early-1990s. In fact, in the textile industry, the difference between the min and max ERP is about 120 percentage points, which constitutes quite an aggressive policy reform over a relatively short time-frame.

5.4 Trade liberalization and firm productivity: A few results

Before proceeding with our discussion of the dynamics of firm-level productivity before and after Colombia's trade policy reforms, a couple of brief comments on the productivity estimates themselves might be in order. First, the ACF, GNR, and LP identification strategies often give rise to substantially different productivity

estimates. In table 5.4, we report industry-level Spearman correlations for the three alternative measures. We find that there is a fair amount of positive comovement between the ACF and GNR estimates, whereas their respective pairwise correlations with the LP measure tend to be much lower and even negative at times. This is quite a remarkable outcome, particularly since the Spearman correlation coefficients are intended to reflect the extent to which the ranking of firms' productivity remains consistent across the three identification strategies. An additional point that is worth mentioning is that the ACF, GNR, and LP productivity estimates do not exhibit the same amount of dispersion. Given that the first of these is based on a value-added model, it displays more heterogeneity than the latter two which arise from a gross output specification of the production function. Industry-level coefficients of variation for the three estimates are reported in table 5.3. Now that we are aware of these differences in characteristics across the ACF, GNR, and LP measures of productivity, we are ready to move on to a summary of the main results of this chapter. In Section 5.4.1, we discuss the output of a simple quantile regression model in which the dependent and independent variables are firm-level productivity and an industry-level indicator of trade policy, respectively. In Section 5.4.2, we apply the methodological framework of Melitz and Polanec (2015) and quantify the relative contributions of incumbent firms, new entrants, and exiting firms to industry-level productivity growth as Colombia switched from a protectionist to a more liberalized trade policy regime during the 1980s.

5.4.1 Quantile regression coefficient estimates

In table 5.5, we report coefficient estimates for a number of different specifications of a quantile regression model in which the log of firm productivity and the 4-digit industry-level import tariff are the dependent and explanatory variables, respectively. For each of the LP, ACF, and GNR measures of productivity, we estimate

three equations - one that includes an industry dummy, another that includes both an industry and a time dummy, and finally, one where consideration is limited to industries that are categorized as "import-competing". This latter categorization has been applied in previous studies that examine the empirical link between trade policy and productivity, most notably in Pavcnik (2002), who defines a 4-digit industry as import-competing if the ratio of imports to total output exceeds a particular threshold. The author experiments with different cutoff values and finds that her results remain fairly consistent when the ratio lies between 0.10 and 0.25. In her final analysis, she settles on 0.15, which is the value that we use here as well.

As explained in Cameron and Trivedi (2005), the quantile regression coefficient estimates $\hat{\boldsymbol{\beta}}_q$ are obtained by minimizing the objective function below, which is an extension of the framework originally proposed by Koenker and Bassett (1978):

$$Q(\boldsymbol{\beta}_q) = \sum_{i,t:\omega_{it} \geq \mathbf{x}'_{it}\boldsymbol{\beta}_q}^{N,T} q|\omega_{it} - \mathbf{x}'_{it}\boldsymbol{\beta}_q| + \sum_{i,t:\omega_{it} < \mathbf{x}'_{it}\boldsymbol{\beta}_q}^{N,T} (1-q)|\omega_{it} - \mathbf{x}'_{it}\boldsymbol{\beta}_q|, \tag{5.17}$$

where ω_{it} denotes the log of firm i's productivity in period t, $\mathbf{x}_{it}$ is a vector of explanatory variables that includes the trade policy indicator and the industry/year dummies, and the q subscript in $\boldsymbol{\beta}_q$ reflects the fact that the coefficient estimates vary by quantile. The coefficients in $\hat{\boldsymbol{\beta}}_q$ are therefore interpreted as the estimated change in the q^{th} quantile of log firm productivity ω_{it} when there is a unit change in the explanatory variables in $\mathbf{x}_{it}$.[8] In the present study, we estimate $\boldsymbol{\beta}_q$ using the approach in (5.17) for $q = 0.10, 0.25, 0.5, 0.75, 0.90$. Hence, this empirical strategy allows us to shed light on the statistical association between indicators of trade protectionism such as import tariffs or the ERP and the *overall distribution of firm-level productivity* which, as mentioned earlier, is one of the most important testable

[8]Note that the tariff and ERP variables are expressed in decimal form, so the estimated change in the q^{th} quantile of log firm productivity given a percentage-point change in the tariff/ERP would be given by the coefficient $\hat{\beta}_q$ divided by 100.

implications of the theoretical framework of Melitz (2003).

Three key findings in table 5.5 are worth emphasizing. First the sign of the quantile regression coefficient estimates displays a fair amount of sensitivity to the manner in which the production function has been specified. In columns 1-6 where the log of the LP and ACF productivity estimates are the dependent variables, most of the coefficient estimates - especially those that correspond to the median, upper quartile, and top decile - are negative. However, when the Cobb-Douglas specification of LP and ACF is replaced by the nonparametric framework of GNR in columns 7-9, we often observe a *positive* association between import tariffs and firm productivity. Second, regardless of whether the LP, ACF, or GNR estimation procedures are employed, there is more of an adverse association between the tariff rate and firm-level productivity in the right tail than in the left tail of the distribution of firms. Evidence of this phenomenon can be seen in pretty much every single column of table 5.5, where the regression coefficient estimates tend to be in decline as one moves downward from the row that corresponds to the 0.10^{th} quantile to the one that corresponds to the 0.90^{th} quantile. Hence, much of the productivity growth that was experienced during Colombia's era of trade liberalization appears to have taken place among firms who were already among the most efficient in their respective industries. Third, when we restrict our attention to import-competing industries, which constitues about one-quarter of the sample, we find some evidence of a negative relationship between import tariffs and the top three quantiles of the GNR productivity estimates. We also observe moderate increases in the coefficient estimates for the bottom decile, lower quartile, and median of the LP and ACF measures of firm productivity.

Next, we re-estimate the nine quantile regression models that have just been discussed, but where the effective rate of protection (ERP) now serves as the indicator of trade policy. The results are reported in table 5.6. In this instance, we observe

a similar pattern to what was noted in the previous paragraph. The negative association between the ERP and productivity becomes more pronounced as one moves closer to the right-tail of the distribution of firms. Thus, during the period of trade liberalization from the mid-1980s to the early-1990s, the most productive firms seem to have made the greatest efficiency gains, and this finding holds across each of the LP, ACF, or GNR identification strategies. In addition, there is an interesting point of divergence between the regression models that use the tariff rate and the ERP, respectively, as the explanatory variable reflecting the trade policy regime. While the former yields positive and statistically significant coefficient estimates for the lower-half of the distribution of GNR productivity, the latter gives rise to coefficient estimates that are either negative or quite small in magnitude relative to their standard errors (or both). This suggests that of the three different measures of productivity that are considered in this chapter, the one that relies on the most flexible (i.e. nonparametric) specification of the production function exhibits a more ambiguous statistical relationship with the indicators of trade protectionism than the measures arising from a more traditional linearized Cobb-Douglas functional form.

5.4.2 Decomposition of aggregate productivity changes

Melitz and Polanec (2015) build on previous work by Olley and Pakes (1996) and outline a framework for the decomposition of industry-level productivity changes into the respective contributions of surviving firms, new entrants, and exiting firms. In the present context, let $t \in \{H, L\}$ denote a time period that is characterized by either a high (H) or a low (L) tariff regime, and let $j \in \{S, X, E\}$ denote the group to which firm i belongs, namely either survivors (S), exiters (X), or entrants (E). Note that in the Colombian manufacturing data, the high-tariff period generally precedes the low-tariff period, and hence the exiting firms and new entrants only appear in the sample in periods H and L, respectively. Let ω_{ijt} denote firm i's productivity

and let s_{ijt} represent its share of industry-level output under tariff regime t, where the subscript j serves to indicate that firm i belongs to group j. Thus, group j's share of aggregate output in period t is given by $s_{jt} = \sum_i s_{ijt}$ and its aggregate productivity is computed as $\Phi_{jt} = \sum_i \frac{s_{ijt}}{s_{jt}} \omega_{ijt}$. Melitz and Polanec (2015) then show that aggregate industry-level productivity under the tariff regimes H and L can be written as:

$$\Phi_H = s_{SH}\Phi_{SH} + s_{XH}\Phi_{XH}$$

$$\Phi_L = s_{SL}\Phi_{SL} + s_{EL}\Phi_{EL}.$$

This gives rise to the following decomposition of the change in aggretate productivity $\Delta\Phi$ when an industry's trade policy regime switches from H to L:

$$\begin{aligned} \Delta\Phi &= (\Phi_{SL} - \Phi_{SH}) + s_{EL}(\Phi_{EL} - \Phi_{SL}) + s_{XH}(\Phi_{SH} - \Phi_{XH}) \\ &= \Delta\bar{\omega}_S + \Delta\text{cov}_S + s_{EL}(\Phi_{EL} - \Phi_{SL}) + s_{XH}(\Phi_{SH} - \Phi_{XH}), \end{aligned} \tag{5.18}$$

where $\Delta\bar{\omega}_S$ denotes the change in the mean productivity of surviving firms, Δcov_S denotes the change in the covariance of surviving firms' productivity and their share of total output, and $s_{EL}(\Phi_{EL} - \Phi_{SL})$ and $s_{XH}(\Phi_{SH} - \Phi_{XH})$ respectively capture the effects of entry of more productive firms and exit of less productive firms in the intervening period between the high tariff and low tariff regimes. Thus, Melitz and Polanec's (2015) framework makes it possible to test some of the theoretical predictions that are found in Melitz (2003) and quantify the relative importance of four distinct channels through which trade liberalization is believed to affect industry-level productivity.

Tables 5.7 through 5.10 contain the raw results of the decomposition exercise that has been performed using the LP, ACF, and GNR measures of productivity. All of the reported values have been normalized by setting $\Phi_H = 1$ for each in-

dustry. Two different samples - respectively comprising the years in which import tariffs and the effective rate of protection attain their max and min values - are once again used for the analysis.[9] Given that the large volume of the raw results makes them somewhat difficult to interpret, we provide a simplified summary of some of the key findings in tables 5.11 through 5.13. To begin, in table 5.11, we report the frequency with which the aggregate and decomposed estimates of firm productivity growth exhibit a positive sign, as might be predicted by modern trade theory. Regardless of whether the tariff rate or the ERP is used as the indicator of protectionism, the first column shows that aggregate ACF productivity experiences positive change with the greatest frequency; in this instance, the sign of $\Delta\Phi$ is greater than zero in nearly three-quarters of the industries that appear in the sample. On the other hand, when the 3-digit industry-level change in ERP is considered, aggregate LP productivity growth is positive only half of the time. In regard to the decomposed growth estimates, we find that efficiency gains among surviving firms ($\Delta\bar{\omega}_S$), efficient reallocation of market share among incumbents (Δcov_S), and the exit of inefficient firms ($s_{XH}(\Phi_{SH} - \Phi_{XH})$) tend to play a more important role than the entry of productive firms into the market ($s_{EL}(\Phi_{EL} - \Phi_{SL})$). While the latter is characterized by a positive sign in less than half of the industries in our sample, its magnitude is generally very small, and hence we conclude that it rarely makes any noteworthy contribution to industry-level productivity growth.

Tables 5.12 and 5.13 shed light on the consistency of the results of the decomposition exercise across the LP, ACF, and GNR productivity measures. The former includes Spearman correlations of the three estimates of each of the growth components in (5.18), while the latter reports on a pairwise basis the frequency with which they display the same expected positive sign. Here, we observe one of this chapter's

[9]The max of both tariffs and the ERP tends to be observed in the mid-1980s, while the min tends to be observed in either the late 1980s (tariffs) or the early 1990s (ERP), due to differences in data availability.

more interesting results, namely that there is far less uniformity than might originally have been anticipated in the dynamics of the LP, ACF, and GNR estimates as Colombia shifted from a protectionist to a more liberalized trade policy regime. The Spearman correlations in table 5.12 are quite modest and in some cases, are actually negative. The decomposition procedure gives rise to particularly different outcomes under the LP and GNR identification strategies. Table 5.13 reflects a similar tendency whereby under the very best scenario, the various components of productivity growth only exhibit the same sign across the different measures of productivity in about half of the industries in the sample. Moreover, this finding does not change when we move from the 4-digit industry-level tariff to the 3-digit industry-level ERP as the trade policy indicator in the model. Hence, any judgement about the relative contributions of incumbent firms, exiters, and new entrants to industry-level productivity growth ultimately depends on the underlying specification of the production function. If we wish to evaluate the performance of firms and industries subsequent to trade policy reforms, it is therefore imperative that we keep in mind the sensitivity of the Melitz-Polanec framework in (5.18) to the choice of a Cobb-Douglas functional form vs. a more flexible nonparametric alternative.

5.5 Conclusion

This chapter has applied three commonly-used strategies for identifying production functions and has examined whether a consistent pattern emerges vis-à-vis the dynamics of firm productivity during periods of trade policy reform in the Colombian manufacturing sector. It has found that the statistical association between productivity and both the nominal and effective rate of protection is rather sensitive to the chosen production function estimation procedure. Switching from a value-added to a gross output model, or from a Cobb-Douglas "control function" framework to a

more flexible nonparametric specification tends to alter the results of our quantile regression model and of the productivity growth decomposition exercise that we perform for a number of manufacturing industries. This raises questions about whether previous empirical findings in the productivity and trade literature are robust to alternative specifications of the production function. Extending our analysis to other firm-level datasets offers interesting possibilities for future research.

References

Ackerberg, D., K. Caves, and G. Frazer (2006): "Structural identification of production functions," Mpra paper, University Library of Munich, Germany.

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics: Methods and Applications*, New York, NY: Cambridge University Press.

Fernandes, A. (2007): "Trade policy, trade volumes and plant-level productivity in Colombian manufacturing industries," *Journal of International Economics*, 71, 52–71.

Gandhi, A., S. Navarro, and D. Rivers (2016): "On the identification of production functions: How heterogeneous is productivity?" Manuscript.

Hu, A. G. and Z. Liu (2014): "Trade liberalization and firm productivity: Evidence from Chinese manufacturing industries," *Review of International Economics*, 22, 488–512.

Koenker, R. and G. Bassett (1978): "Regression quantiles," *Econometrica*, 46, 33–50.

Levinsohn, J. and A. Petrin (2003): "Estimating production functions using inputs to control for unobservables," *Review of Economic Studies*, 70, 317–341.

Melitz, M. J. (2003): "The impact of trade on intra-industry reallocations and aggregate industry productivity," *Econometrica*, 71, 1695–1725.

Melitz, M. J. and S. Polanec (2015): "Dynamic Olley-Pakes productivity decomposition with entry and exit," *RAND Journal of Economics*, 46, 362–375.

Olley, S. G. and A. Pakes (1996): "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 64, 1263–97.

Pavcnik, N. (2002): "Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants," *Review of Economic Studies*, 69, 245–276.

Roberts, M. and J. Tybout (1997): "The decision to export in Colombia: An empirical model of entry with sunk costs," *American Economic Review*, 87, 545–64.

Schor, A. (2004): "Heterogeneous productivity response to tariff reduction: Evidence from Brazilian manufacturing firms," *Journal of Development Economics*, 75, 373–396.

Topalova, P. and A. Khandelwal (2011): "Trade liberalization and firm productivity: The case of India," *Review of Economics and Statistics*, 93, 995–1009.

Tybout, J. R. and M. D. Westbrook (1995): "Trade liberalization and the dimensions of efficiency change in Mexican manufacturing industries," *Journal of International Economics*, 39, 53–78.

5.6 Appendix

ISIC	min tariff	max tariff	ISIC	min tariff	max tariff
3111	0.265	0.411	3513	0.266	0.398
3112	0.298	0.470	3521	0.288	0.463
3113	0.413	0.653	3522	0.092	0.139
3114	0.285	0.446	3523	0.367	0.637
3115	0.170	0.401	3529	0.214	0.343
3116	0.260	0.382	3551	0.181	0.274
3117	0.375	0.620	3559	0.397	0.585
3118	0.169	0.289	3560	0.490	0.812
3119	0.000	0.633	3620	0.269	0.430
3121	0.259	0.415	3691	0.218	0.351
3122	0.090	0.110	3692	0.136	0.224
3131	0.568	0.930	3699	0.280	0.453
3132	0.488	0.798	3710	0.185	0.288
3133	0.325	0.475	3811	0.337	0.559
3134	0.400	0.660	3812	0.400	0.797
3211	0.403	0.815	3813	0.252	0.449
3212	0.655	1.224	3819	0.326	0.550
3213	0.672	1.331	3821	0.089	0.205
3214	0.700	1.255	3822	0.059	0.189
3215	0.450	0.721	3823	0.152	0.270
3219	0.460	0.806	3824	0.162	0.271
3220	0.657	1.217	3825	0.220	0.476
3231	0.191	0.336	3829	0.236	0.417
3232	0.425	0.425	3831	0.254	0.474
3233	0.464	0.775	3832	0.226	0.329
3240	0.564	0.934	3833	0.365	1.005
3311	0.382	0.604	3839	0.284	0.468
3312	0.445	0.735	3841	0.173	0.287
3319	0.358	0.592	3842	0.197	0.496
3320	0.400	0.823	3843	0.367	0.578
3411	0.223	0.369	3844	0.404	0.772
3412	0.392	0.647	3845	0.101	0.198
3419	0.308	0.482	3849	0.371	0.613
3420	0.362	0.511	3851	0.196	0.314
3511	0.180	0.290	3852	0.208	0.329
3512	0.054	0.128	Pooled	0.000	1.331

Table 5.1: Import tariffs in the Colombian manufacturing sector 1981-1988 (4-digit ISIC).

ISIC	min ERP	max ERP	ISIC	min ERP	max ERP
311	0.791	1.470	352	0.250	0.413
313	0.574	1.349	355	0.536	1.004
321	0.826	2.033	356	0.712	1.467
322	0.734	1.900	362	0.360	0.561
323	0.441	0.990	369	0.383	0.625
324	0.821	1.674	371	0.242	0.395
331	0.649	1.182	381	0.585	0.988
332	0.565	1.371	382	0.163	0.372
341	0.415	0.668	383	0.370	0.815
342	0.360	0.595	384	0.504	1.058
351	0.208	0.378	385	0.224	0.428

Table 5.2: Effective rate of protection in the Colombian manufacturing sector 1981-1991 (3-digit ISIC).

ISIC	LP	ACF	GNR	ISIC	LP	ACF	GNR
311	0.651	1.280	0.258	352	0.369	0.739	0.203
313	0.051	0.840	0.407	355	0.281	0.564	0.342
321	1.745	0.489	0.199	356	0.333	0.581	0.148
322	0.352	0.586	0.147	362	0.154	0.464	0.230
323	0.431	0.562	0.123	369	0.204	0.829	0.254
324	0.449	0.485	0.114	371	0.180	0.659	0.297
331	0.144	0.376	0.148	381	0.101	0.529	0.157
332	0.147	0.308	0.137	382	1.393	0.771	0.219
341	0.235	2.249	0.222	383	0.098	0.583	0.178
342	0.175	0.458	0.144	384	0.114	0.602	0.258
351	0.164	0.736	0.234	385	0.369	0.590	0.169

Table 5.3: Coefficient of variation of LP, ACF, and GNR productivity estimates by industry.

	LP	ACF	GNR		LP	ACF	GNR
311				**352**			
LP	1			LP	1		
ACF	0.355	1		ACF	0.181	1	
GNR	-0.402	-0.794	1	GNR	0.093	0.875	1
313				**355**			
LP	1			LP	1		
ACF	0.341	1		ACF	0.557	1	
GNR	0.338	0.885	1	GNR	0.517	0.891	1
321				**356**			
LP	1			LP	1		
ACF	0.639	1		ACF	-0.058	1	
GNR	0.384	0.802	1	GNR	-0.303	0.858	1
322				**362**			
LP	1			LP	1		
ACF	0.306	1		ACF	0.083	1	
GNR	0.249	0.887	1	GNR	-0.321	0.750	1
323				**369**			
LP	1			LP	1		
ACF	0.624	1		ACF	0.209	1	
GNR	0.478	0.779	1	GNR	0.033	0.829	1
324				**371**			
LP	1			LP	1		
ACF	0.289	1		ACF	-0.003	1	
GNR	-0.070	0.724	1	GNR	-0.369	0.809	1
331				**381**			
LP	1			LP	1		
ACF	-0.418	1		ACF	0.197	1	
GNR	-0.705	0.819	1	GNR	-0.008	0.885	1
332				**382**			
LP	1			LP	1		
ACF	0.284	1		ACF	0.023	1	
GNR	-0.203	0.743	1	GNR	-0.055	0.944	1
341				**383**			
LP	1			LP	1		
ACF	0.328	1		ACF	0.248	1	
GNR	0.134	0.881	1	GNR	0.255	0.938	1
342				**384**			
LP	1			LP	1		
ACF	0.176	1		ACF	0.170	1	
GNR	-0.073	0.802	1	GNR	0.129	0.844	1
351				**385**			
LP	1			LP	1		
ACF	0.267	1		ACF	-0.416	1	
GNR	-0.103	0.755	1	GNR	-0.787	0.701	1

Table 5.4: Spearman correlations of productivity estimates by 3-digit industry.

	LP			ACF			GNR		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
q_{10}	-0.0048	-0.0067	-0.0014	-0.0153	0.0440	0.1258	0.1240	0.2352	0.0789
	(0.0014)	(0.0016)	(0.0047)	(0.0185)	(0.0224)	(0.0637)	(0.0191)	(0.0238)	(0.0637)
q_{25}	-0.0112	-0.0150	0.0031	-0.0798	-0.0302	0.0638	0.0896	0.1934	0.0454
	(0.0025)	(0.0031)	(0.0083)	(0.0155)	(0.0187)	(0.0522)	(0.0134)	(0.0168)	(0.0402)
q_{50}	-0.0341	-0.0477	0.0049	-0.1675	-0.1485	-0.0323	0.0459	0.1211	-0.0077
	(0.0044)	(0.0053)	(0.0135)	(0.0166)	(0.0204)	(0.0529)	(0.0090)	(0.0108)	(0.0319)
q_{75}	-0.0730	-0.0964	-0.0427	-0.2402	-0.2249	-0.2303	0.0130	0.0600	-0.0775
	(0.0074)	(0.0091)	(0.0210)	(0.0204)	(0.0261)	(0.0722)	(0.0068)	(0.0084)	(0.0300)
q_{90}	-0.1112	-0.1419	-0.1410	-0.3535	-0.3740	-0.2699	0.0097	0.0420	-0.0723
	(0.0151)	(0.0181)	(0.0414)	(0.0318)	(0.0379)	(0.1139)	(0.0075)	(0.0093)	(0.0367)
3-digit ISIC effects	yes	yes	yes	yes	yes	yes	yes	yes	yes
year effects		yes			yes			yes	
import-competing			yes			yes			yes
N	38,297	38,297	8,478	38,297	38,297	8,478	38,297	38,297	8,478

Table 5.5: Quantile regression output where tariff rate is the trade policy indicator (standard errors in parentheses). Dependent variable is log of plant-level productivity. Data available for years 1981, 1983-1988.

	LP			ACF			GNR		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
q_{10}	-0.0018	-0.0011	-0.0005	-0.0157	-0.0057	0.0212	-0.0027	0.0399	-0.0610
	(0.0009)	(0.0013)	(0.0032)	(0.0107)	(0.0152)	(0.0409)	(0.0132)	(0.0176)	(0.0376)
q_{25}	-0.0052	-0.0055	0.0085	-0.0228	-0.0122	0.0128	-0.0064	0.0266	0.0017
	(0.0016)	(0.0022)	(0.0062)	(0.0101)	(0.0143)	(0.0334)	(0.0090)	(0.0124)	(0.0309)
q_{50}	-0.0113	-0.0158	0.0020	-0.0459	-0.0506	-0.0210	-0.0117	-0.0008	0.0056
	(0.0027)	(0.0038)	(0.0102)	(0.0110)	(0.0152)	(0.0358)	(0.0058)	(0.0081)	(0.0215)
q_{75}	-0.0241	-0.0335	-0.0138	-0.1032	-0.1110	-0.0973	-0.0126	-0.0006	-0.0375
	(0.0046)	(0.0064)	(0.0152)	(0.0144)	(0.0193)	(0.0455)	(0.0042)	(0.0061)	(0.0197)
q_{90}	-0.0415	-0.0627	0.0104	-0.1362	-0.1317	-0.1612	-0.0159	-0.0075	-0.0174
	(0.0097)	(0.0137)	(0.0275)	(0.0230)	(0.0315)	(0.0747)	(0.0044)	(0.0061)	(0.0225)
3-digit ISIC effects	yes	yes	yes	yes	yes	yes	yes	yes	yes
year effects		yes			yes			yes	
import-competing			yes			yes			yes
N	27,177	27,177	6,287	27,177	27,177	6,287	27,177	27,177	6,287

Table 5.6: Quantile regression output where ERP is the trade policy indicator (standard errors in parentheses). Dependent variable is log of plant-level productivity. Data available for years 1981, 1984, 1985, 1990, 1991.

ISIC	$\Delta\Phi$	$\Delta\bar{\omega}_S$	Δcov_S	$s_{XH}\left(\Phi_{SH}-\Phi_{XH}\right)$	$s_{EL}\left(\Phi_{EL}-\Phi_{SL}\right)$
3111	0.109	0.275	-0.162	-0.001	-0.003
3112	0.187	0.060	0.123	0.012	-0.008
3113	-0.048	0.000	-0.052	0.006	-0.001
3114	0.007	-0.055	0.031	0.011	0.019
3115	0.069	0.109	-0.043	0.006	-0.004
3116	0.109	-0.039	0.165	0.012	-0.028
3117	0.020	0.007	0.009	0.008	-0.005
3118	0.017	-0.017	0.035	-0.001	-0.001
3119	0.080	0.039	0.049	0.014	-0.023
3121	0.054	0.021	0.029	0.005	-0.001
3122	0.115	0.044	0.077	0.007	-0.013
3131	0.033	0.005	0.026	0.002	0.000
3132	0.001	0.018	-0.016	0.001	-0.001
3133	0.016	0.003	0.013	0.000	0.000
3134	0.014	-0.001	0.016	0.000	-0.001
3211	0.050	0.019	0.004	0.013	0.014
3212	0.039	-0.024	0.044	0.010	0.009
3213	-0.003	-0.008	0.009	0.001	-0.006
3214	-0.028	-0.034	0.009	0.000	-0.004
3215	-0.023	-0.009	-0.014	0.000	0.000
3219	-0.145	0.007	-0.146	0.005	-0.011
3220	-0.003	-0.01	0.011	0.006	-0.010
3231	0.053	0.068	-0.018	0.001	0.002
3233	0.092	0.361	-0.257	-0.007	-0.005
3240	0.045	-0.048	0.09	0.017	-0.015
3311	-0.012	-0.002	-0.023	0.034	-0.020
3312	-0.053	-0.004	-0.025	0.000	-0.023
3319	-0.007	-0.038	0.027	-0.004	0.008
3320	0.001	-0.011	0.034	-0.005	-0.016
3411	-0.048	-0.005	-0.039	-0.005	0.000
3412	0.064	0.030	0.040	0.005	-0.011
3419	-0.010	0.008	0.003	-0.015	-0.006
3420	0.037	-0.004	0.042	-0.003	0.003
3511	0.030	0.015	-0.006	0.001	0.020
3512	0.026	0.060	-0.048	0.004	0.010
3513	-0.006	0.028	-0.048	0.004	0.010
3521	0.113	0.033	0.092	0.005	-0.018
3522	-0.017	0.008	-0.030	-0.001	0.006
3523	-0.032	0.012	0.019	-0.002	-0.061
3529	0.037	0.029	0.010	0.001	-0.003
3551	-0.052	0.019	-0.071	0.003	-0.003
3559	0.000	0.004	-0.008	0.013	-0.009
3560	0.07	-0.003	0.035	0.001	0.038
3620	-0.062	-0.014	-0.051	-0.001	0.004
3691	-0.003	-0.013	0.005	-0.001	0.006
3692	0.044	0.041	0.038	-0.030	-0.005
3699	0.106	0.021	0.145	-0.074	0.015
3710	0.033	-0.015	0.045	0.003	0.000
3811	0.004	0.001	0.002	0.003	-0.002
3812	-0.024	0.009	-0.028	-0.002	-0.003
3813	-0.011	0.001	-0.014	0.006	-0.004
3819	-0.024	0.002	-0.027	-0.001	0.002
3821	-0.014	-0.003	-0.012	0.000	0.000
3822	-0.005	-0.071	0.071	0.002	-0.007
3823	-0.535	-0.006	-0.004	-0.532	0.006
3824	0.001	0.005	0.006	-0.004	-0.006
3825	0.038	0.002	0.041	0.007	-0.013
3829	0.023	0.017	0.013	-0.004	-0.003
3831	0.019	0.019	0.002	0.003	-0.005
3832	0.022	0.011	0.009	0.007	-0.006
3833	0.016	0.014	-0.008	0.006	0.004
3839	0.013	-0.002	0.015	0.002	-0.002
3841	0.004	0.003	-0.045	0.002	0.045
3842	-0.014	-0.043	0.008	0.028	-0.006
3843	0.005	0.017	-0.011	0.003	-0.004
3844	0.018	-0.022	0.021	0.039	-0.020
3845	-0.130	-0.055	-0.074	0.000	-0.001
3849	0.018	0.043	-0.021	0.000	-0.004
3851	0.028	-0.025	0.045	-0.004	0.011
3852	-0.032	-0.170	0.136	-0.004	0.006

Table 5.7: Melitz-Polanec decomposition of LP productivity growth following tariff cut.

ISIC	$\Delta\Phi$	$\Delta\bar{\omega}_S$	Δcov_S	$s_{XH}\left(\Phi_{SH}-\Phi_{XH}\right)$	$s_{EL}\left(\Phi_{EL}-\Phi_{SL}\right)$
3111	0.136	0.053	0.058	-0.024	0.049
3112	0.095	0.074	0.021	0.029	-0.029
3113	-0.251	-0.020	-0.243	0.010	0.002
3114	-0.230	-0.077	-0.004	-0.130	-0.018
3115	-0.152	-0.079	-0.073	0.004	-0.004
3116	0.125	0.032	0.117	0.015	-0.039
3117	-0.030	0.074	-0.109	0.005	0.000
3118	-0.113	-0.015	-0.091	-0.002	-0.005
3119	0.118	0.087	0.128	-0.038	-0.059
3121	0.110	0.103	0.022	0.004	-0.020
3122	0.040	0.029	0.022	-0.015	0.003
3131	0.233	0.046	0.176	0.012	-0.001
3132	-0.108	-0.212	0.104	0.012	-0.012
3133	0.303	0.314	-0.010	-0.001	0.000
3134	-0.049	-0.105	0.048	0.011	-0.002
3211	0.165	0.045	0.023	0.019	0.078
3212	0.265	0.000	0.324	0.015	-0.073
3213	0.007	-0.008	0.021	0.004	-0.008
3214	-0.093	-0.108	0.015	0.000	0.000
3215	0.050	0.064	-0.014	0.000	0.000
3219	0.013	0.094	-0.066	0.006	-0.021
3220	0.133	-0.017	0.179	0.009	-0.037
3231	-0.074	0.123	-0.168	-0.033	0.005
3233	0.015	-0.031	0.026	-0.039	0.059
3240	0.015	0.023	-0.028	0.018	0.002
3311	-0.093	0.014	-0.037	-0.037	-0.033
3312	0.341	0.182	0.055	0.000	0.103
3319	0.034	0.047	-0.003	0.008	-0.018
3320	0.288	0.077	0.338	-0.012	-0.115
3411	0.043	0.256	-0.199	-0.010	-0.004
3412	0.096	0.046	0.067	0.019	-0.035
3419	0.005	-0.137	0.234	-0.083	-0.008
3420	0.001	0.065	-0.071	0.003	0.005
3511	0.090	0.025	-0.032	-0.027	0.124
3512	0.077	0.228	-0.198	0.012	0.034
3513	0.381	0.205	0.220	-0.078	0.034
3521	0.395	0.096	0.404	-0.040	-0.065
3522	0.440	0.020	0.053	-0.014	0.382
3523	0.093	0.071	0.063	0.004	-0.046
3529	0.175	0.208	-0.001	-0.009	-0.023
3551	0.187	0.063	0.127	0.006	-0.009
3559	0.122	0.044	0.076	0.013	-0.011
3560	0.134	0.014	0.039	0.005	0.077
3620	0.067	-0.022	0.012	-0.002	0.078
3691	0.012	0.021	-0.031	0.021	0.001
3692	0.061	-0.050	0.118	0.004	-0.010
3699	0.067	0.032	0.066	-0.119	0.089
3710	0.163	0.155	0.001	0.011	-0.003
3811	0.177	0.135	0.040	-0.002	0.004
3812	1.924	0.074	2.248	-0.027	-0.370
3813	0.296	0.129	0.076	-0.023	0.114
3819	0.287	0.077	0.213	0.019	-0.022
3821	-0.244	-0.126	-0.118	0.000	0.000
3822	-0.116	-0.066	-0.014	0.004	-0.040
3823	-0.017	0.050	-0.066	-0.094	0.093
3824	0.104	0.073	0.010	0.001	0.020
3825	0.504	0.406	0.317	-0.114	-0.105
3829	0.289	0.174	0.171	0.116	-0.171
3831	0.443	0.228	0.213	-0.007	0.009
3832	0.113	0.041	0.065	0.015	-0.008
3833	0.927	0.054	0.096	0.018	0.760
3839	0.212	0.096	0.118	0.002	-0.004
3841	0.402	0.192	0.472	-0.004	-0.258
3842	-0.326	-0.346	-0.009	0.019	0.009
3843	-0.044	0.034	-0.075	0.007	-0.011
3844	-0.053	-0.080	0.050	0.242	-0.267
3845	-0.265	0.114	-0.375	0.000	-0.004
3849	0.063	0.130	-0.074	0.000	0.007
3851	0.918	0.038	0.250	0.004	0.626
3852	-0.080	-0.011	-0.078	0.000	0.009

Table 5.8: Melitz-Polanec decomposition of ACF productivity growth following tariff cut.

ISIC	$\Delta\Phi$	$\Delta\bar{\omega}_S$	Δcov_S	$s_{XH}\left(\Phi_{SH}-\Phi_{XH}\right)$	$s_{EL}\left(\Phi_{EL}-\Phi_{SL}\right)$
3111	-0.009	-0.028	0.039	0.016	-0.036
3112	-0.017	-0.013	-0.003	-0.014	0.014
3113	-0.108	-0.011	-0.108	0.012	0.000
3114	-0.080	-0.080	-0.125	0.127	-0.002
3115	0.027	-0.011	0.032	0.002	0.003
3116	0.022	-0.029	0.051	-0.015	0.014
3117	0.068	-0.007	0.084	-0.010	0.001
3118	-0.077	-0.017	-0.062	0.001	0.001
3119	-0.051	-0.006	-0.046	-0.008	0.009
3121	-0.043	-0.027	-0.025	0.001	0.008
3122	-0.060	-0.051	-0.014	0.003	0.002
3131	0.182	0.052	0.119	0.012	-0.001
3132	-0.077	-0.163	0.066	0.007	0.013
3133	0.032	-0.072	0.103	0.000	0.000
3134	0.042	-0.051	0.082	0.014	-0.003
3211	0.031	0.000	0.021	0.016	-0.006
3212	0.010	-0.045	0.023	0.011	0.020
3213	-0.007	0.001	0.000	-0.004	-0.004
3214	-0.034	-0.015	-0.018	0.000	-0.002
3215	0.008	0.012	-0.004	0.000	0.000
3219	-0.010	0.050	-0.057	0.001	-0.005
3220	-0.010	-0.006	-0.005	0.003	-0.001
3231	0.135	0.072	0.057	0.001	0.005
3233	-0.014	-0.026	0.028	-0.023	0.008
3240	-0.035	0.03	-0.074	-0.003	0.011
3311	-0.028	-0.002	0.001	-0.009	-0.018
3312	0.147	0.071	0.030	0.000	0.046
3319	0.032	0.021	0.014	0.008	-0.011
3320	0.048	0.025	0.031	-0.002	-0.005
3411	0.030	0.056	-0.024	-0.004	0.002
3412	0.078	0.059	0.020	0.006	-0.008
3419	0.108	0.024	0.110	-0.022	-0.004
3420	0.068	0.064	0.016	-0.022	0.010
3511	-0.026	-0.009	-0.001	-0.005	-0.010
3512	-0.002	-0.029	0.029	0.003	-0.005
3513	0.071	0.024	0.056	-0.011	0.003
3521	0.130	-0.006	0.147	0.009	-0.020
3522	0.017	0.003	0.010	-0.007	0.011
3523	-0.017	0.038	0.037	0.003	-0.095
3529	0.048	0.057	0.000	0.002	-0.011
3551	0.277	0.073	0.208	0.003	-0.007
3559	0.081	0.018	0.068	-0.002	-0.003
3560	0.004	0.009	0.003	-0.006	-0.001
3620	0.245	0.064	0.143	-0.011	0.050
3691	0.032	-0.015	0.029	0.023	-0.006
3692	0.203	0.116	0.042	0.050	-0.0053
3699	0.102	0.026	0.098	-0.033	0.010
3710	0.090	0.084	0.005	0.006	-0.006
3811	0.079	0.063	0.017	-0.001	0.000
3812	0.033	0.017	0.029	-0.012	-0.002
3813	0.135	0.050	0.069	-0.011	0.028
3819	0.019	-0.001	0.010	0.008	0.001
3821	-0.050	-0.048	-0.002	0.000	0.000
3822	-0.026	-0.024	0.018	0.003	-0.023
3823	-0.013	0.021	-0.033	-0.067	0.066
3824	0.069	0.045	0.016	0.000	0.009
3825	-0.006	-0.008	0.045	-0.050	0.006
3829	0.098	0.063	0.059	0.021	-0.045
3831	0.097	0.026	0.054	0.001	0.017
3832	0.015	-0.003	0.011	0.008	-0.001
3833	0.093	0.019	0.080	0.009	-0.014
3839	0.049	0.043	0.014	0.004	-0.012
3841	0.287	0.120	0.421	-0.004	-0.250
3842	-0.275	-0.267	-0.011	0.010	-0.007
3843	0.076	0.033	0.046	0.000	-0.003
3844	0.021	-0.043	0.050	-0.057	0.071
3845	-0.341	-0.036	-0.302	0.000	-0.003
3849	0.024	0.002	0.034	0.000	-0.012
3851	-0.027	-0.010	0.024	0.005	-0.046
3852	-0.059	-0.004	-0.048	0.002	-0.009

Table 5.9: Melitz-Polanec decomposition of GNR productivity growth following tariff cut.

ISIC	$\Delta\Phi$	$\Delta\bar{\omega}_S$	Δcov_S	$s_{XH}(\Phi_{SH}-\Phi_{XH})$	$s_{EL}(\Phi_{EL}-\Phi_{SL})$
			Levinsohn-Petrin		
311	0.078	0.024	0.059	-0.001	-0.003
313	-0.019	0.000	-0.030	0.008	0.003
321	-0.013	0.018	-0.008	-0.014	-0.008
322	0.108	0.021	0.068	0.041	-0.022
323	0.052	0.252	-0.206	0.011	-0.006
324	0.127	-0.013	0.137	0.017	-0.014
331	0.077	0.001	0.082	0.023	-0.029
332	0.013	0.032	0.002	-0.003	-0.019
341	0.016	0.011	0.006	-0.001	0.000
342	0.046	-0.017	0.054	-0.002	0.011
351	-0.024	0.000	-0.034	0.008	0.001
352	0.010	0.016	0.021	-0.004	-0.023
355	-0.064	0.086	-0.152	0.009	-0.007
356	0.023	0.018	0.012	-0.006	-0.001
362	-0.073	0.000	-0.078	0.002	0.004
369	0.069	0.023	0.087	-0.047	0.006
371	-0.131	-0.064	-0.082	0.011	0.003
381	-0.017	0.005	-0.02	-0.001	0.000
382	-0.075	0.010	0.020	-0.085	-0.02
383	-0.008	-0.005	-0.009	0.006	0.000
384	-0.006	0.007	-0.018	0.007	-0.002
385	0.202	0.036	0.141	-0.003	0.029
			Ackerberg-Caves-Frazer		
311	-0.358	-0.128	-0.246	-0.019	0.036
313	-0.108	-0.02	-0.200	0.075	0.037
321	0.117	0.071	0.095	-0.020	-0.029
322	0.350	0.067	0.182	0.142	-0.041
323	-0.069	0.058	-0.127	-0.010	0.01
324	1.186	0.105	1.396	0.047	-0.361
331	0.674	0.085	0.640	0.068	-0.118
332	0.101	0.056	0.067	0.014	-0.035
341	0.055	0.002	0.058	0.000	-0.005
342	0.047	0.073	-0.023	-0.018	0.014
351	0.003	-0.025	0.026	-0.028	0.029
352	0.017	0.089	-0.030	-0.006	-0.035
355	-0.170	-0.018	-0.160	0.017	-0.01
356	-0.039	-0.056	0.013	0.011	-0.007
362	-0.074	-0.08	-0.069	0.008	0.068
369	0.013	-0.013	0.068	-0.077	0.035
371	0.089	-0.010	0.094	0.011	-0.005
381	0.289	-0.028	0.332	0.003	-0.019
382	0.224	0.232	0.156	-0.072	-0.092
383	0.021	-0.027	0.037	0.020	-0.01
384	0.051	-0.040	0.103	-0.029	0.017
385	0.176	-0.042	0.063	0.018	0.137
			Gandhi-Navarro-Rivers		
311	0.018	-0.001	0.000	0.024	-0.005
313	-0.079	-0.012	-0.153	0.079	0.007
321	0.017	0.003	0.020	0.006	-0.012
322	0.019	-0.001	0.013	0.015	-0.008
323	0.101	0.010	0.091	0.000	-0.001
324	0.084	-0.003	0.068	0.007	0.012
331	0.092	0.023	0.063	0.017	-0.011
332	0.032	-0.010	0.038	0.008	-0.004
341	0.033	0.037	-0.010	0.000	0.006
342	0.067	0.068	-0.023	-0.019	0.041
351	-0.028	-0.017	-0.004	-0.002	-0.005
352	0.000	0.034	0.003	-0.001	-0.036
355	-0.204	0.051	-0.262	0.008	0.000
356	-0.013	-0.028	0.011	0.003	0.000
362	0.172	0.013	0.119	-0.007	0.047
369	0.146	0.023	0.109	0.006	0.007
371	0.058	0.034	0.027	0.003	-0.005
381	0.013	-0.025	0.029	-0.001	0.01
382	0.089	0.110	0.001	-0.016	-0.006
383	0.010	-0.021	0.031	0.005	-0.005
384	-0.056	-0.015	-0.053	-0.005	0.017
385	-0.080	-0.061	-0.015	0.008	-0.012

Table 5.10: Melitz-Polanec decomposition of productivity growth following ERP cut.

	$\Delta\Phi$	$\Delta\bar{\omega}_S$	Δcov_S	$s_{XH}\left(\Phi_{SH}-\Phi_{XH}\right)$	$s_{EL}\left(\Phi_{EL}-\Phi_{SL}\right)$
$\Delta Tariff$					
LP	0.614	0.571	0.600	0.600	0.314
ACF	0.743	0.729	0.629	0.543	0.400
GNR	0.614	0.529	0.700	0.514	0.400
ΔERP					
LP	0.545	0.727	0.545	0.500	0.455
ACF	0.727	0.455	0.682	0.545	0.409
GNR	0.682	0.500	0.682	0.682	0.455

Table 5.11: Proportion of industry-level productivity changes that have expected positive sign.

	$\Delta Tariff$				ΔERP		
$\Delta\Phi$							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.221	1		ACF	0.260	1	
GNR	0.102	0.523	1	GNR	0.133	0.221	1
$\Delta\bar{\omega}_S$							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.293	1		ACF	-0.112	1	
GNR	0.147	0.435	1	GNR	-0.032	0.596	1
Δcov_S							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.198	1		ACF	0.390	1	
GNR	-0.110	0.482	1	GNR	0.065	0.328	1
$s_{XH}\left(\Phi_{SH}-\Phi_{XH}\right)$							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.312	1		ACF	0.571	1	
GNR	0.072	0.443	1	GNR	0.294	0.609	1
$s_{EL}\left(\Phi_{EL}-\Phi_{SL}\right)$							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.468	1		ACF	0.819	1	
GNR	-0.019	0.208	1	GNR	0.444	0.320	1

Table 5.12: Spearman correlation of component-wise LP, ACF, and GNR productivity growth.

$\Delta Tariff$				ΔERP			
$\Delta\Phi$							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.493	1		ACF	0.409	1	
GNR	0.423	0.535	1	GNR	0.409	0.545	1
$\Delta\bar{\omega}_S$							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.500	1		ACF	0.364	1	
GNR	0.343	0.471	1	GNR	0.364	0.318	1
Δcov_S							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.429	1		ACF	0.409	1	
GNR	0.414	0.529	1	GNR	0.409	0.500	1
$s_{XH}\left(\Phi_{SH}-\Phi_{XH}\right)$							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.400	1		ACF	0.364	1	
GNR	0.371	0.400	1	GNR	0.364	0.455	1
$s_{EL}\left(\Phi_{EL}-\Phi_{SL}\right)$							
	LP	ACF	GNR		LP	ACF	GNR
LP	1			LP	1		
ACF	0.225	1		ACF	0.273	1	
GNR	0.155	0.239	1	GNR	0.273	0.227	1

Table 5.13: Frequency with which the pairwise LP, ACF, and GNR productivity growth components have the same expected positive sign.

Chapter 6

Conclusion

In this dissertation, I have covered a variety of semiparametric and nonparametric methods for the econometric analysis of firm-level production data. I have considered three distinct approaches, namely deterministic frontier analysis, stochastic frontier analysis, and proxy function methods of controlling for unobserved productivity, and in each case, I have offered new methodological insights that should be appealing to practitioners. I have highlighted the importance of two robustness criteria that an estimator ought to satisfy in the context of modelling the production technology of firms; first, the estimation procedure should apply to a number of different specifications of the function that relates input combinations and output quantities, or the distribution that characterizes firms' productive efficiency. The traditional parametric log-linearized Cobb-Douglas framework is generally unsatisfactory in this regard because it does not allow for much flexibility in how factor elasticities are modelled. Similarly, the parametric stochastic frontier model for panel data is inadequate whenever it is built on erroneous assumptions about the density of firm-level inefficiency. In contrast, the semiparametric and nonparametric approaches that have been delineated in Chapters 2 and 4 of this thesis are rather flexible in terms of the functional form and distributional assumptions that they

rely upon, and in this sense, they satisfy the first robustness criterion.

In regard to the second robustness criterion, Chapter 3 of this dissertation has shown that extreme-valued observations can be a very serious concern in the context of deterministic production frontier estimation. When data envelopment methods are being applied, a single outlier that has arisen due to measurement error can severely distort the estimated boundary of a production set. Thus, it is unsurprising that the literature has sought to address the challenge of defining a frontier estimator that does not explicitly allow for random noise or measurement error (hence the "deterministic" label), but that is nevertheless robust to outliers. The approach that I have proposed improves upon existing robust estimation procedures insofar as it makes use of hierarchical clustering to determine an optimal amount of data trimming; this results in as many non-extreme-valued observations as possible getting enveloped by the frontier, while outliers end up being dropped from the analysis. Chapter 3 has also acknowledged that in certain industries, the output of firms is more realistically expressed as a count rather than a continuous quantity and hence, in instances such as these, the existing robust frontier estimation framework needs to be modified accordingly. A notable example of this phenomenon is found in the R & D sector, where the production of proprietary technology is measured in terms of firm-level patent counts.

The methodological considerations that have been put forward in this thesis ought to prove useful to those who are engaged in applied research and/or policy analysis that involves firm-level data. Chapter 5 provided a clear example of how a simple change in model specification can alter one's assessment of the success or failure of a policy reform that is intended to promote productivity growth at the firm-level. It is therefore important to consider a number of different empirical strategies when modelling input-output relationships and the productive efficiency of firms, and the various approaches that have been outlined here can be beneficial

in this regard. Although it might not have been immediately obvious to the reader, this dissertation has also served to highlight just a few of the many tasks that can be performed in the R statistical computing environment when applying kernel methods to the analysis of microdata. All of the smooth estimators that were described in Chapters 2 through 4 were implemented using the `np` and `crs` packages in R, and the applied researcher who is interested in semiparametric and nonparametric econometric analysis will find that these offer a more comprehensive range of functions than what is currently available in some of the proprietary statistical software packages that are popular in the social science community.