Heterogeneity in Firms: A Proxy Variable Approach for Quantile Production Functions

Justin Doty*and Suyong Song†

January 21, 2021

Abstract

We propose a new approach to estimate firm-level production functions in which output elasticities are heterogeneous across the firm-size distribution. This paper extends the proxy variable approach for estimating production functions to the conditional quantiles of firm production. Production function parameters are identified by conditional quantile restrictions and estimated using the implied unconditional sample moment restrictions. We show that this method allows us to capture heterogeneity in output elasticities along the firm-size distribution that would not be estimated in conditional mean models. We provide small-sample evidence in a Monte Carlo study to show that this approach is robust compared to other production function estimators. The method is applied to firm and plant-level manufacturing data from the US, Chile, and Colombia.

Keywords: Production functions, Heterogeneous elasticity, Nonlinear quantile regression

JEL Classification: C14, C36, D24

1 Introduction

Production function estimation is an ongoing and historical empirical research topic that links firm's input to output decisions. Identification of the output elasticities and consequently the distribution of firm-level productivity is constrained by endogeneity issues. This is because productivity is unobserved by the econometrician, but observed by the firm when making input decisions.

A popular approach to address this issue is to introduce a proxy variable such as investment, made popular by Olley and Pakes (1996) or an intermediate material input using Levinsohn and Petrin (2003) or Ackerberg *et al.* (2015). These proxies are a function of a state variable such as capital and the unobserved productivity components. Under certain assumptions, this demand function is strictly increasing in its scalar unobserved productivity component. Inverting this

*Department of Economics, University of Iowa, S321 Pappajohn Business Building, 21 E Market St, Iowa City, IA 52242. Email: justin-doty@uiowa.edu

†Department of Economics and Finance, University of Iowa, W360 Pappajohn Business Building, 21 E Market St, Iowa City, IA 52242. Email: suyong-song@uiowa.edu

demand function controls for unobserved productivity and the production function parameters can be estimated with a simple two-stage estimator.

While these methods have been useful in identifying the production function parameters and recovering consistent estimates of total factor productivity (TFP) resulting estimates may be biased if there is additional heterogeneity in production technology across firms. Thus, allowing for heterogeneous coefficients is one possible way to capture these differences. The literature on heterogeneous production functions is small relative to the empirical research using the homogeneous coefficient model, even though many empirical studies have found firm's heterogeneous behavior and decision.[1] This is because estimating the homogeneous coefficient model by itself is very difficult due to the issue of unobserved productivity.

In our approach we allow firm heterogeneity in production technology beyond Hick's neutral productivity shock to be driven by the rank of the unobserved production shock, η_{it}. We simultaneously extend the proxy variable approach to this framework in order to control for the part of production unobservables that are correlated with inputs. Since applying the quantile regression requires non-smooth criterion function, it is not straightforward to estimate the production functions by allowing for endogenous inputs and their heterogeneous coefficients. We are not aware of any published paper which takes into account for the endogeneity issue of production functions in the conventional quantile regression framework. We fill the gap in this paper by proposing an easy-to-implement estimator.

We show through simulation, that our proposed two-step estimator performs relatively well to the most current control function approach of Levinsohn and Petrin (2003) and is successful in capturing heterogeneous output elasticities along the conditional distribution of firm's output. In our empirical application, we consider several popular firm and plant-level manufacturing datasets and compare our estimator to control function approaches. We show that heterogeneity in these estimates implies differences in other features of the production function, such as capital intensity and TFP growth over time.

The rest of the paper is organized as follows. Section 2 reviews prior approaches for production function estimation and the literature on panel data quantile regression. Section 3 introduces the econometric model and the proposed estimator. Section 4 presents finite-sample behaviors of the estimator via Monte Carlo experiments and Section 5 applies this estimator to US, Chilean, and Colombian manufacturing datasets. Section 6 concludes with directions for future research.

[1]Some notable examples are Kasahara, Schrimpf and Suzuki (2017), Balat, Brambilla and Sasaki (2018), Li and Sasaki (2017) and Dermirer (2020) to name of few. Also Gandhi *et al.* (2020) who estimate a nonparametric production function and obtain heterogeneous estimates by construction.

2 Literature Review

2.1 Production Function Estimation

We briefly review the LP (2003) procedure for estimating a *value-added* production function (in logs) [2,3].

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \eta_{it}. \tag{1}$$

where y_{it} denotes value-added output, l_{it} denotes labor input for firm i at time t, k_{it} denotes capital input, ω_{it} is unobserved productivity and η_{it} denotes an iid shock to production.

To control for the correlation between ω_{it} and inputs k_{it} and l_{it}. LP introduce an intermediate input demand defined as [4]

$$m_{it} = m_t(k_{it}, \omega_{it}) \tag{2}$$

where the function f is strictly increasing in ω_{it} for all k_{it}. Productivity can then be expressed as

$$\omega_{it} = m_t^{-1}(k_{it}, m_{it}). \tag{3}$$

Substituting into the production function

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + m_t^{-1}(k_{it}, m_{it}) + \eta_{it} = \beta_l l_{it} + \Phi(k_{it}, m_{it}) + \eta_{it} \tag{4}$$

An estimate for β_l and $\Phi_t(k_{it}, m_{it})$ can be obtained by the following first stage moment restriction

$$\mathbb{E}[y_{it} - \beta_l l_{it} - \Phi_t(k_{it}, m_{it}) | \mathcal{I}_{it}] = 0 \tag{5}$$

where $\mathcal{I}_{it}$ denotes the firm's information at time t. A linear approximation can be used, in which case estimates can be obtained from a simple linear regression.

A second stage moment restriction identifies the coefficient on capital. Assume that productivity follows an auto-regressive process

$$\omega_{it} = \mathbb{E}[\omega_{it} | \omega_{it-1}] + \xi_{it} = g(\omega_{it-1}) + \xi_{it} \tag{6}$$

where ξ_{it} denotes an innovation to productivity and satisfies $\mathbb{E}[\xi_{it} | \mathcal{I}_{it-1}] = 0$.

[2] We consider a value-added production function here to be consistent with the model we introduce in Section 3 for reasons which we will discuss in the corresponding section

[3] We drop the constant β_0 since it is not separately identified from ω_{it} without a location normalization

[4] In the original paper of Levinsohn and Petrin (2003) they consider multiple intermediate inputs such as energy, fuels, and materials as potential proxies. We focus on material inputs as the proxy.

Then, the production function parameters can be estimated from the moment restrictions

$$\begin{aligned}\mathbb{E}[\xi_{it} + \eta_{it} | \mathcal{I}_{it-1}] = & \\ & \mathbb{E}[\tilde{y}_{it} - \beta_k k_{it} \\ & \quad - g(\hat{\Phi}_{t-1}(k_{it-1}, m_{it-1}) - \beta_k k_{it-1}) | \mathcal{I}_{it-1}] = 0,\end{aligned} \tag{7}$$

where $\tilde{y}_{it} = y_{it} - \hat{\beta}_l l_{it}$ and $\hat{\Phi}$ denotes estimates from the first stage. LP proceed by using instruments from $\mathcal{I}_{it-1}$ and minimize a Generalized Method of Moments (GMM) criterion function. Standard errors are obtained using a bootstrap procedure since the two-step nature of this estimators complicates asymptotic inference.

2.2 Production Functions and Quantile Regression

Connecting variation in the random error, η_{it}, to differences in a firm's final output decisions is not straightforward in the standard production function model. We briefly review a subfield of production function estimation that facilitates a more natural interpretation; production frontier models. We discuss limitations of these applications and return to our interpretation in Section 3.

A (stochastic) frontier (SFA) model of production proposed by Aigner *et al.* (1977) introduces statistical error into a frontier model. Frontier models assume firms firms deviate from an optimal frontier of production. The SFA model is typically written as

$$y_i = f(x_i, \beta) + \varepsilon_i, \tag{8}$$

where $\varepsilon_i = \eta_i - u_i$, x_i are inputs to production and β are the parameters. The error term η_i denotes the statistical noise in the model such as measurement error and u_{it} represents one-sided deviations from the production frontier. Estimates of β are typically obtained using maximum likelihood which requires strong distributional assumptions on the error terms. Estimation of the efficient frontier is then a conditional mean estimator rather than a maximal value estimator as noted by Bernini *et al.* (2004). They suggest quantile regression could then be used to estimate the highest percentiles of the conditional output distribution as it relates to the stochastic frontier, however, a theoretical difficulty is then choosing which quantile corresponds to the frontier. A more detailed derivation of a quantile representation of the frontier was introduced by Aragon *et al.* (2005) in a nonparametric model which requires inversion of a conditional empirical CDF. Since the purpose of this paper is not to compare the advantages and disadvantages of production frontier models and ours we leave this discussion for future studies and acknowledge the theoretical challenges of quantile frontier models.

There are two main challenges of implementing a quantile regression framework to the standard production function model. Firstly, as we alluded to earlier in this section, if we maintain a structural interpretation of firm production, reduced-form linear quantile regression models are

likely not sufficient in linking a firm's output choice as a function of the error term η_{it} as we elaborate in the next section. Secondly, addressing the endogeneity of ω_{it} using traditional panel data methods have challenges specific towards the production function literature and quantile models.

Regarding the second point, quantile panel data models allow for flexible interactions between unobserved heterogeneity and the quantiles of the conditional response function. Some well known approaches assume a time-invariant fixed effect such as Koenker (2004), Lamarche (2010), Canay (2011) which acts as a pure location shifter of the conditional quantile function. This approach may have two main disadvantages. First, assuming the unobservable is time-invariant is restrictive and Griliches and Hausman (1986) has shown to leads to low estimates of β_k. Secondly, the fixed effects of these models are incidental parameters so as the sample size grows, so does the number of parameters that need to be estimated which makes it computationally costly. An alternative to fixed effect estimation is to model the unobserved heterogeneity as a projection onto the observables plus a disturbance in the spirit of Chamberlain (1984). Abrevaya and Dahl (2008) adopt this approach with a linear data generating process for birth outcomes and linking it to its quantile function to estimate the effect of birth inputs over the birth-weight distribution. This approach is further developed by Bache *et al.* (2012). One downside of this approach is that it is difficult to describe the behavior of the conditional quantile function as it depends on the joint distribution of unobservables in the response function and the random effect.

Another alternative is to make use of valid instruments if they are available. The conventional argument for using input prices p_{it}^k and p_{it}^l as instruments is that they must be uncorrelated with the error term $\omega_{it} + \eta_{it}$ and correlated with input choices for capital and labor. Then one could use two-stage least squares to obtain consistent estimates of β_k and β_l. This idea can be extended to quantile-IV models such as Chernozhukov and Hansen (2005). In their identification arguments, one would need to strengthen assumptions to conditional independence as well as monotonicity of a quantile structural function (QSF) in $U_{it} = \omega_{it} + \eta_{it}$. Then if one writes the QSF for the production function as $y_{it} = Q(k_{it}, l_{it}, U_{it})$ where $\tau \in (0, 1]$ denotes the quantile index, the model is identified from a quantile type moment restriction

$$P[y_{it} \leq Q(k_{it}, l_{it}, \tau) | k_{it}, l_{it}, p_{it}^k, p_{it}^l] = \tau \tag{9}$$

while this identification argument is used in our model, we do not use the estimation procedure for reasons explained in the next section. One of these reasons are that input prices may not have enough variation across firms and exogeneity can be violated if they capture input quality differences as argued by Griliches and Hausman (1986).

3 A Random Coefficient Production Function

We specify a *value-added* production function as a random coefficient model:

$$y_{it} = \beta_k(\eta_{it})k_{it} + \beta_l(\eta_{it})l_{it} + \omega_{it} \tag{10}$$

A value-added specification in equation (10) is non-trivial. Value-added production functions are common in the empirical literature, however the objects recovered from a value-added model such as the output elasticities and TFP can only be mapped to its gross-output counterpart under specific structural production functions. Since the production shock enters (10) non-separably, it is difficult to recover gross-output objects from value-added due to the presence of ex-post shocks. Therefore, in our empirical application we interpret the elasticities and TFP with some caution.

One advantage of the value-added approach is that the rank of η_{it} can now be interpreted as the rank of firm-size as measured by value-added which accounts for firm size from the value of output created through capital and labor. There may be other reasons to measure firm-size by other measures such as gross-output (sales) or employment which is sometimes refined as employee weighted average of number of employees (Kumar *et al.*, 1999). The value-added approach also avoids the non-identification results of Gandhi *et al.* (2020). We leave the connection between this value-added production function and its possibly underlying gross-output production function as future work.

The variables in equation (10) have the same interpretation as the ones we introduced in the LP model. The only difference here is that we allow the output elasticities to be functionally dependent on the production shock η_{it} while productivity still maintains its additive separability. [5]

A special case of (10) is the location scale model,

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + (\mu_k k_{it} + \mu_l l_{it})\eta_{it} \tag{11}$$

Which implies that the τth conditional quantile of y_{it} is given by

$$Q_{y_{it}}(\tau|\mathcal{I}_{it}) = \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + (\mu_k k_{it} + \mu_l l_{it})F^{-1}(\tau) \tag{12}$$

where $F^{-1}(\tau)$ is the quantile function of production shocks η_{it}.

The formulation of (11) is not new to the production function literature. The assumption that input choices can impact firm's production beyond the conditional mean has important con-

[5] More specifically, productivity is only a location shifter of the conditional output distribution. We cannot allow $\omega_{it} = \omega_{it}(\eta_{it})$ since this would violate the scalar unobservability assumption of our proxy variable

sequences for firm's attitude towards production risk. A volume of literature that originated in the late 1970's challenged the standard stochastic specifications of production functions (Just and Pope, 1978, 1979) by considering a specification that allowed firm's inputs to both increase or decrease the marginal variability of final output. The most common application of those models are in the agricultural industry where the variance on the yield of harvested crops could be increased by adverse weather or decreased by pesticide usage. Since manufacturing businesses tend to operate in a more controlled environment, risk is less prevalent in these industries so the conditional variance of η_{it} may be smaller. A general quantile model such as the one specified in (10) can be seen as an extension of the higher-order moment estimation of risk initiated by Antle (1983). However, it can also be seen purely as an econometric specification issue as we are unaware of any tests that could distinguish between higher order moment production risk and misspecification. We choose the latter interpretation for our model.

We note that under quantile preferences a firm who maximizes the τ level of utility of profits could explain heterogeneity in the output distribution. Unlike risk-neutral firms, firms could have a utility function that is represented by preferences of the firm manager(s) who decides the optimal expenditure on inputs. Different managers may have different preferences for risk. Quantile utility maximization is not a new concept. A short list of papers have considered quantile utility maximization such as Manski (1988), Rostek (2009), Chambers (2007), and Bhattacharya (2009). Dynamic input choices such as investment are much more difficult to solve using the quantile utility framework and the reader can refer to de Castro and Galvao (2017) for a treatment of dynamic quantile utility models. As far as we know, the quantile utility framework has not been applied to firm decision problems and a more thorough treatment of such is outside the scope of this paper.

3.1 Identification

3.1.1 Production Function in Levinsohn and Petrin (2003)

Returning to our misspecification interpretation in Equation (10), we follow LP in the usual set of assumptions on timing of input choices and scalar unobservability.

Assumption 3.1

(a) The production function $y_{it} = f_t(k_{it}, l_{it}, \omega_{it}, \eta_{it})$ is strictly increasing in η_{it}

(b) The firm's information set at time t includes current and past productivity shocks $\{\omega_{it}\}_{t=0}^{t}$, but does not include past productivity shocks $\{\omega_{it}\}_{t=t+1}^{\infty}$. η_{it} is independent of $\mathcal{I}_{it}$

(c) Firm's productivity shocks evolve according to a first-order Markov process

$$\omega_{it} = g(\omega_{it-1}) + \xi_{it} \tag{13}$$

where the iid productivity innovations ξ_{it} satisfy $\mathbb{E}[\xi_{it}|\mathcal{I}_{it-1}] = 0$

(d) Firms accumulate capital according to

$$K_{it} = \kappa_t(I_{it-1}, K_{it-1}). \tag{14}$$

where K_{it-1} and I_{it-1} denote previous period capital and investment

(e) Firm's intermediate input demand function is given by $m_{it} = m_t(k_{it}, \omega_{it})$

(f) The intermediate input demand function $m_t(k_{it}, \omega_{it})$ is strictly increasing in ω_{it}

Given these Assumption (3.1)(e, f), we invert intermediate input demand $\omega_{it} = m^{-1}(k_{it}, m_{it})$ and substitute into the production function. We treat m_t^{-1} as a nonparametric function (k_{it}, m_{it}). We then have:

$$y_{it} = \beta_k(\eta_{it})k_{it} + \beta_l(\eta_{it})l_{it} + m_t^{-1}(k_{it}, m_{it}) = \beta_l(\eta_{it})l_{it} + \Phi(k_{it}, m_{it}, \eta_{it}) \tag{15}$$

Using Assumption (3.1)(a, b) we have the following identification condition for the first stage:

$$P\big(y_{it} \leq \beta_l(\tau)l_{it} + \Phi(k_{it}, m_{it}; \tau)\big|\mathcal{I}_{it}\big) = \tau \tag{16}$$

The equation in (15) is a semiparametric partially linear quantile regression model which can be consistently estimated using approaches proposed by Lee (2003), Koekner *et al.* (1994), or Chen and Pouzo (2009). We can then plug these estimates back into the production function and apply Assumption (3.1)(c) to obtain:

$$y_{it} = \beta_k(\eta_{it})k_{it} + \hat{\beta}_l(\tau)l_{it} + g(\hat{\Phi}(k_{it-1}, m_{it-1}; \tau) - \beta_k(\eta_{it})k_{it-1}) + \xi_{it} \tag{17}$$

The main challenge of the identification result in the second stage is that it is not true in general that

$$P\big(y_{it} \leq \beta_k(\tau)k_{it} + \hat{\beta}_l(\tau)l_{it} + g(\hat{\Phi}(k_{it-1}, m_{it-1}; \tau) - \beta_k(\tau)k_{it-1}) + \xi_{it}|\mathcal{I}_{it-1}\big) = \tau \tag{18}$$

The lack of identification in using the above equation is related to the identification issues of quantile panel data models using correlated random effects. For example, Canay (2011) notes that the conditional behavior of (18) depends on the joint distribution of η_{it} and ξ_{it} which makes identification of $(g, \beta_k(\tau))$ in the second stage problematic. Identification issues related to quantile panel data models using conditional quantile restrictions are studied by Rosen (2012) who develops conditions for both set and point identification. Cai *et al.* (2018) note that an equation like (17) can be interpreted as measurement error in the dependent variable of a quantile regression model where ξ_{it} is the measurement error in y_{it}. This type of model is studied by Hausman *et al.* (2019) who use

a sieve MLE approach for estimating the density of the measurement error. Both aforementioned papers require varying degrees of distributional assumptions on the error term which we would like to avoid in our model. An alternative would be to use the entropic integration method of Schennach (2014) and integrate out ξ_{it} via simulation. However, this approach would introduce additional nuisance parameters that would need to be estimated.

We can avoid distributional assumptions and nuisance parameters by framing the identification condition in the ξ_{it} component, similar to Ackerberg *et al.* (2015) by concentrating out the constant $\beta_0(\tau)$ and g. For a hypothetical guess of $\beta_k(\tau)$, say $\tilde{\beta}_k(\tau)$ we can write for a fixed $\tau \in (0, 1]$

$$\beta_0(\tau) + \omega_{it} = y_{it} - \hat{\beta}_l(\tau) l_{it} - \tilde{\beta}_k(\tau) k_{it} = \hat{\Phi}(k_{it}, m_{it}; \tau) - \tilde{\beta}_k(\tau) k_{it} \tag{19}$$

We can rewrite the AR(1) productivity process as

$$\hat{\Phi}(k_{it}, m_{it}; \tau) - \tilde{\beta}_k(\tau) k_{it} = \beta_0(\tau) + g(\hat{\Phi}(k_{it-1}, m_{it-1}; \tau) - \tilde{\beta}_k(\tau) k_{it-1}) + \xi_{it} \tag{20}$$

and note that the implied residuals satisfy

$$\mathbb{E}[\xi_{it}(\tilde{\beta}_k(\tau)) | \mathcal{I}_{it-1}] = 0 \tag{21}$$

which can be estimated by standard GMM procedures for a fixed $\tau \in (0, 1]$

This estimator is then a simple extension of the control function approach where the first stage parameters are estimated using quantile regression techniques. The location-shift assumption on the productivity process allows us to identify and estimate the parameters in the second stage while concentrating out additional parameters. Some interesting extensions could include additional heterogeneity in the productivity process such as a location-scale model for productivity which could be estimated using He (1997) or a random-coefficient model where second stage identification would rely on a conditional quantile restriction and estimated using Kaplan and Sun (2016).

3.1.2 Production Function in Ackerberg *et al.* (2015)

In the Ackerberg *et al.* (2015) setting, the intermediate input demand $m_{it} = m_t(k_{it}, l_{it}, \omega_{it})$ is conditional on the labor input. This allows a more flexible timing assumption on when labor is chosen by the firm relative to the other inputs. Labor can have dynamic implications and be partially or fully realized before productivity ω_{it}. In this setting, the labor elasticity β_l cannot be identified in the first stage as in the LP approach. The first stage equation is then:

$$y_{it} = \beta_k(\eta_{it}) k_{it} + \beta_l(\eta_{it}) l_{it} + m_t^{-1}(k_{it}, l_{it}, m_{it}) = \Phi(k_{it}, l_{it}, m_{it}, \eta_{it}), \tag{22}$$

where we have used monotonicity of the intermediate input demand function in ω_{it} to control for unobserved productivity. The nonparametric function $\Phi(\cdot;\tau)$ can be identified using Assumption (3.1) (a) and (b)

$$P\big(y_{it} \leq \Phi(k_{it}, l_{it}, m_{it}; \tau)\big|\mathcal{I}_{it}\big) = \tau \tag{23}$$

which suggests that the functional $\Phi(\cdot,\tau)$ can be estimated by nonparametric quantile methods such local linear or polynomial regression (Chaudhuri, 1991a,b), smoothing splines (Koekner *et al.*, 1994), or more general sieve based estimation (Chen and Pouzo, 2012).
As before, we concentrate out the constant $\beta_0(\tau)$ and g. For a hypothetical guess of $(\beta_l(\tau), \beta_k(\tau))$ we can write

$$\beta_0(\tau) + \omega_{it} = y_{it} - \tilde{\beta}_l(\tau)l_{it} - \tilde{\beta}_k(\tau)k_{it} = \hat{\Phi}(k_{it}, l_{it}, m_{it}; \tau) - \tilde{\beta}_k(\tau)k_{it} - \tilde{\beta}_l(\tau)l_{it} \tag{24}$$

We can rewrite the AR(1) productivity process as

$$\begin{aligned}\hat{\Phi}(k_{it}, l_{it}, m_{it}; \tau) - \tilde{\beta}_k(\tau)k_{it} - \tilde{\beta}_l(\tau)l_{it} \\ = \beta_0(\tau) + g(\hat{\Phi}(k_{it-1}, l_{it-1}, m_{it-1}; \tau) - \tilde{\beta}_k(\tau)k_{it-1} - \tilde{\beta}_l(\tau)l_{it-1}) + \xi_{it}\end{aligned} \tag{25}$$

and note that the implied residuals satisfy

$$\mathbb{E}[\xi_{it}(\tilde{\beta}_k(\tau), \tilde{\beta}_l(\tau))|\mathcal{I}_{it-1}] = 0 \tag{26}$$

which can be estimated by standard GMM procedures for a fixed $\tau \in (0, 1]$

3.2 Estimation

We discuss how to estimate the LP production function in two stages.

First Stage

Recall, in the first stage we have the identification condition

$$P\big(y_{it} \leq \beta_l(\tau)l_{it} + \Phi(k_{it}, m_{it}; \tau)\big|\mathcal{I}_{it}\big) = \tau \tag{27}$$

which yields

$$\mathbb{E}\big[\mathbb{1}\{y_{it} \leq \beta_l(\tau)l_{it} + \Phi(k_{it}, m_{it}; \tau)\} - \tau\big|\mathcal{I}_{it}\big] = 0 \tag{28}$$

Similar to Olley and Pakes (1996), $\Phi(\cdot;\tau)$ can be approximated by a flexible polynomial so that estimates $\hat{\beta}_l(\tau)$ and $\hat{\Phi}(\cdot;\tau)$ can be obtained from a polynomial quantile regression. A more complete model of $\Phi(\cdot;\tau)$ can be obtained using a finite-dimensional sieve and estimated using a minimum distance criterion function. We briefly introduce this approach.

To fix notation, let $x_{it} = (k_{it}, m_{it})$ denote the variables in the nonparametric function, Φ. Let $\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi(x_{it})) = \mathbb{1}\{y_{it} - \beta_l(\tau) l_{it} - \Phi(x_{it}) \leq 0\} - \tau$. Rephrasing our first stage identification condition as

$$\mathbb{E}[\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi(x_{it}))|\mathcal{I}_{it}] = 0 \tag{29}$$

we can see that this resembles the semiparametric moment conditions studied by Chen and Pouzo (2009) and Ai and Chen (2012) where the residual function is non-differentiable in (β_l, Φ) due to the indicator function. However, one difference between our model and theirs is that there is no endogeneity in the first stage which simplifies estimation. Let $\alpha = (\beta_l, \Phi)$. The first stage estimates can be found from the following minimization problem

$$(\hat{\beta}_l, \hat{\Phi}) = \underset{\alpha \in (\Theta \times \mathcal{H}_{k(n)})}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\mathbb{E}}[\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi)|\mathcal{I}_{it}]' \hat{\Sigma}_1^{-1} \hat{\mathbb{E}}[\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi)|\mathcal{I}_{it}] \tag{30}$$

where $\Theta \subset \mathbb{R}$ with $\beta_l \in \Theta$ and $\{\mathcal{H}_{k(n)} : k(n) = 1, 2, \dots\}$ is a sequence of approximating finite dimensional linear sieve spaces which becomes dense as $k(n) \to \infty$. In practice one could use a tensor-product linear sieve basis function such as B-splines or polynomials. $\hat{\mathbb{E}}[\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi)|\mathcal{I}_{it}]$ and $\hat{\Sigma}_1$ are nonparametric estimators of $\mathbb{E}[\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi)|\mathcal{I}_{it}]$ and Σ_1 which can be obtained using series LS estimator and $\hat{\Sigma}_1 = \tau(1 - \tau)$. In our simulation study and empirical application we use a 3rd order polynomial with interactions and estimate the first stage parameters using simple weighted linear quantile regression, which we justify in the next section.

Second Stage

Once estimates of $\beta_l(\tau)$ and $\Phi(k_{it}, m_{it}; \tau)$ are obtained, we can estimate the residuals of the productivity innovation shocks for a given $\tilde{\beta}_k(\tau)$

$$\hat{\xi}_{it}(\tilde{\beta}_k(\tau)) = \hat{\Phi}(k_{it}, m_{it}; \tau) - \tilde{\beta}_k(\tau) k_{it} - g(\hat{\Phi}(k_{it-1}, m_{it-1}; \tau) - \tilde{\beta}_k(\tau) k_{it-1}, \rho) \tag{31}$$

where we parameterize the process for productivity g by a finite dimensional parameter vector ρ.[6] To simplify notation in the later section we let $\beta = (\beta_0, \beta_k, \rho)$ and $\hat{\xi}_{it}(\tilde{\beta}_k(\tau)) = \Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi(x_{it}))$. so we can rewrite the moment condition in equation (21) as

$$\mathbb{E}[\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi(x_{it}))|\mathcal{I}_{it-1}] = 0 \tag{32}$$

The capital coefficient is identified using the fact that current capital does not respond to innovation shocks in productivity. In practice we include additional instruments $(k_{it-1}, l_{it-1}, m_{it-1})$ so that

[6] We use a third degree polynomial in ω_{it-1} to estimate g in the empirical application. In practice, one could estimate g nonparametrically, however this complicates the asymptotic results

our model is over-identified and $\beta_k(\tau)$ can be estimated using a GMM criterion function so that $\hat{\beta}$ solves

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, \hat{\mathbb{E}}[\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi(x_{it}))]' \hat{\Sigma}_2 \hat{\mathbb{E}}[\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi(x_{it}))] \tag{33}$$

where $\hat{\mathbb{E}}[\cdot]$ are the sample unconditional moments corresponding to (32) and $\hat{\Sigma}_2$ may be an arbitrary weighting matrix. We discuss how to efficiently choose the weighting matrix in this two-step approach using Ackerberg *et al.* (2014) in the next section.

3.3 Asymptotics

We divide the discussion of asymptotics in two parts. In the first part, we show how the coefficients for the variable inputs identified in the first stage can be efficiently estimated using Ai and Chen (2012). In the second part, we show how the capital coefficient can be efficiently estimated using the appropriate weighting matrix as shown in Ackerberg *et al.* (2014).

First Stage

The main difficulty in establishing asymptotic normality for the labor estimate in the first stage is the non-smoothness of the residual function in equation (29). Chen and Pouzo (2009) extend the results of Ai and Chen (2003) to show that the estimate of the parametric part of their model is not only asymptotically normal, but also establish its semiparametric efficiency bound. In the model without endogeneity, they show that their estimator reaches the same efficiency bound as Lee (2003). Readers can refer to Section 5 of Chen and Pouzo (2009), in particular, Remark 5.1 illustrates how to weaken their assumptions in the case of exogeneity. Let $u_{it} = y_{it} - \beta_l l_{it} - \Phi(x_{it})$. Using Condition 5.6 and 5.7 in Chen and Pouzo (2009)

$$\sqrt{n}(\hat{\beta}_l - \beta_l) \to N(0, V_1^{-1}) \tag{34}$$

where

$$V_1 = \tau(1-\tau)\left\{ \mathbb{E}[f^2_{U|L,X}(0) l_{it}^2] - \mathbb{E}\left[\frac{\mathbb{E}\big[f^2_{U|L,X}(0) l_{it} | X\big]^2}{\mathbb{E}\big[f^2_{U|L,X}(0)|X\big]} \right] \right\}^{-1} \tag{35}$$

A simple consistent estimator for the asymptotic variance can be given by

$$\hat{V}_1 = \tau(1-\tau)\left\{ \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\hat{f}^2_{U|L,X}(0) l_{it}^2 - \frac{\hat{\mathbb{E}}\big[f^2_{U|L,X}(0) l_{it} | X\big]^2}{\hat{\mathbb{E}}\big[f^2_{U|L,X}(0)|X\big]} \right] \right\}^{-1} \tag{36}$$

where $\hat{f}_{U|L,X}(0)$ and $\hat{\mathbb{E}}[\cdot|X]$ are consistent nonparametric estimators of $f_{U|L,X}(0)$ and $\mathbb{E}[\cdot|X]$. An alternative to plugging in these into the estimate of the asymptotic variance, is to use a weighted bootstrap procedure. Chen and Pouzo (2009) show that this algorithm produces consistent esti-

mates the asymptotic distribution of β_l as long as one chooses an i.i.d sample of positive weights denoted by w_{it} which satisfy $\mathbb{E}[w_{it}] = 1$ and $Var[w_{it}] < \infty$ and is independent from the data. In practice we draw weights from a standard exponential distribution for each iteration of the bootstrap procedure and use these to compute a weighted quantile regression estimator of β_l and Φ.

Second Stage

Asymptotic normality of two-stage GMM estimators is well-established in the literature. The second stage estimates satisfy

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_2) \tag{37}$$

where

$$V_2 = (G_2' \Sigma_2 G_2)^{-1} (G_2' \Sigma_2 V_{\Lambda 2} \Sigma_2 G_2)(G_2' \Sigma_2 G_2)^{-1} \tag{38}$$

Here, $G_2 = \frac{\partial \mathbb{E}(\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi(x_{it}))}{\partial \beta}, V_{\Lambda 2} = Var(\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi(x_{it})))$ and Σ_2 is the weighting matrix. The two-step GMM estimator of β is not efficient since the information sets used in the first stage and the second stage are not simultaneously considered. In order to adjust for the variance in estimating the first stage parameters β_l and Φ, we show how Ackerberg *et al.* (2014) can be applied to choosing an optimal weighting matrix that reflects the noise from the first stage estimates and provide the semi-parametric efficiency bound.

The intuition of the Ackerberg *et al.* (2014) approach is that under certain conditions, the original unconditional moments $\mathbb{E}[\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi(x_{it}))]$ can be orthogonalized with respect to the moment condition in equation (29). When the first step is exactly identified from a semiparametric partially linear quantile restriction, this new moment condition is written as:[7]

$$\begin{aligned} \tilde{\Lambda}_2(y_{it}, l_{it}, \beta_l, \beta, \Phi) =& \Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi) + \sum_{t=1}^{T} \left[\left(\frac{\partial \mathbb{E}[\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi) | \mathcal{I}_{it-1}]}{\partial l_{it}} \right.\right. \\ & \left.\left. + \frac{\partial \mathbb{E}[\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi) | \mathcal{I}_{t-1}]}{\partial \Phi_j'} \right) \times \frac{\tau - \mathbb{1}\{y_{it} \leq \beta_l l_{it} - \Phi_t(x_{it})\}}{f_{U|L,X}(0)} \right] \end{aligned} \tag{39}$$

Ackerberg *et al.* (2014) show that the semiparametric efficiency bound for β can be written as the inverse of

$$\tilde{V}_2 = \left(\frac{\partial \mathbb{E}[\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi)]}{\partial \beta} \right)' Var(\tilde{\Lambda}_2(y_{it}, l_{it}, \beta_l, \beta, \Phi))^{-1} \left(\frac{\partial \mathbb{E}[\Lambda_2(y_{it}, l_{it}, \beta_l, \beta, \Phi)]}{\partial \beta} \right) \tag{40}$$

[7] Although the results of Chen and Pouzo (2009) and Ai and Chen (2012) do not require the first step to be exactly identified, we require it in order to derive the semiparametric efficiency bound for the second step estimates

so that this bound can be achieved by choosing a weighting matrix $\Sigma_2 = Var(\tilde{\Lambda}_2(y_{it}, l_{it}, \beta_l, \beta, \Phi))^{-1}$ which can be consistently estimated by plugging in estimates $\hat{\beta}_l$, $\hat{\Phi}$, and $\hat{f}_{U|L,X}(0)$ from the first stage.

In practice, it is much easier to compute a numerically equivalent estimate of $Var(\tilde{\Lambda}_2(y_{it}, l_{it}, \beta_l, \beta, \Phi))^{-1}$ as shown by Ackerberg *et al.* (2012) which is outlined by Ackerberg *et al.* (2014) in Section 3.3 of their paper.

4 Monte Carlo Experiments

We use a location-scale version of Levinsohn and Petrin (2003) and replicate Ackerberg *et al.* (2015) simulations sampling 1000 datasets consisting of 1000 firms. We simulate optimal input choices for 100 time periods, using the last 10 periods for estimation.

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + (\gamma_0 + \gamma_k k_{it} + \gamma_l l_{it})\eta_{it} \tag{41}$$

with $\beta_0 = 0, \beta_k = 0.4$ and $\beta_l = 0.6$. The location scale parameters are set as $\gamma_0 = 0, \gamma_k = 0.7$ and $\gamma_l = -0.6$ For each simulation we simulate two DGPs with $\eta_{it} \sim N(0, 0.1)$ and $\eta_{it} \sim Laplace(0, 0.1)$.

To produce consistent estimates of the labor coefficient in the first stage, we do not allow for any wage variation across firms and labor is chosen at time t with perfect information about ω_{it}. However, we add optimization error in labor. An AR(1) process is specified for productivity $\omega_{it} = \rho\omega_{it-1} + \xi_{it}$ where $\rho = 0.7$. The variance of ξ_{it} and initial value ω_{i0} is set so that the standard deviation of ω_{it} is constant over time and equal to 0.3

We compare the LP estimation procedure with our "QLP" two-step procedure under the two different sets of experiments specified earlier. We estimate the model for $\tau \in \{0.1, 0.15, \ldots, 0.85, 0.9\}$ and use current period capital, k_{it} as our instrument so that our model is exactly identified. For the weighting matrix, we use an estimate of the variance covariance matrix of the sample moments. We use a continuously updated GMM procedure such that estimates of $\beta_k(\tau)$ in the second stage are estimated simultaneously with the weighting matrix. We initialize the algorithm at the true value of $\beta_k(\tau)$ however we find that the estimation is robust to reasonable initial values.

We focus on whether our estimate can capture heterogeneity in the output distribution reasonably well compared to the mean estimates from the LP approach. Previous papers such as Olley and Pakes (1996) and Levinsohn and Petrin (2003) already show that the control function approach controls for endogeneity bias from unobserved productivity and perform specification tests for the set of possible control functions such as investment, material inputs, fuels, and energy.

Figure 1 graphs the estimated coefficients for QLP and LP. The black line denotes the QLP estimator and its corresponding 90% confidence interval is the gray shaded area. The solid red line denotes the LP estimator and the dotted red lines are its corresponding 90% confidence interval.

Capital coefficients are the first column of the graph and labor coefficients are the second column. The first row corresponds to DGP 1 with $\eta_{it} \sim N(0, 0.1)$ and the second row corresponds to DGP 2 with $\eta_{it} \sim Laplace(0, 0.1)$. The estimator does reasonable well at capturing heterogeneity outside of the conditional mean estimate for both capital and labor. Not surprisingly, the confidence intervals for capital are wider than the interval for labor due to nature of the multi-step procedure. 2 shows that these estimators perform well in finite sample. The MSE for both estimators is plotted over $\tau \in \{0.1, 0.15, \ldots, 0.85, 0.9\}$ with the black line denoting the QLP estimates and the dotted red line denoting the LP estimates.

Figure 1: QLP estimated coefficients of $\beta_k(\tau)$ and $\beta_l(\tau)$. Dotted line is LP estimator

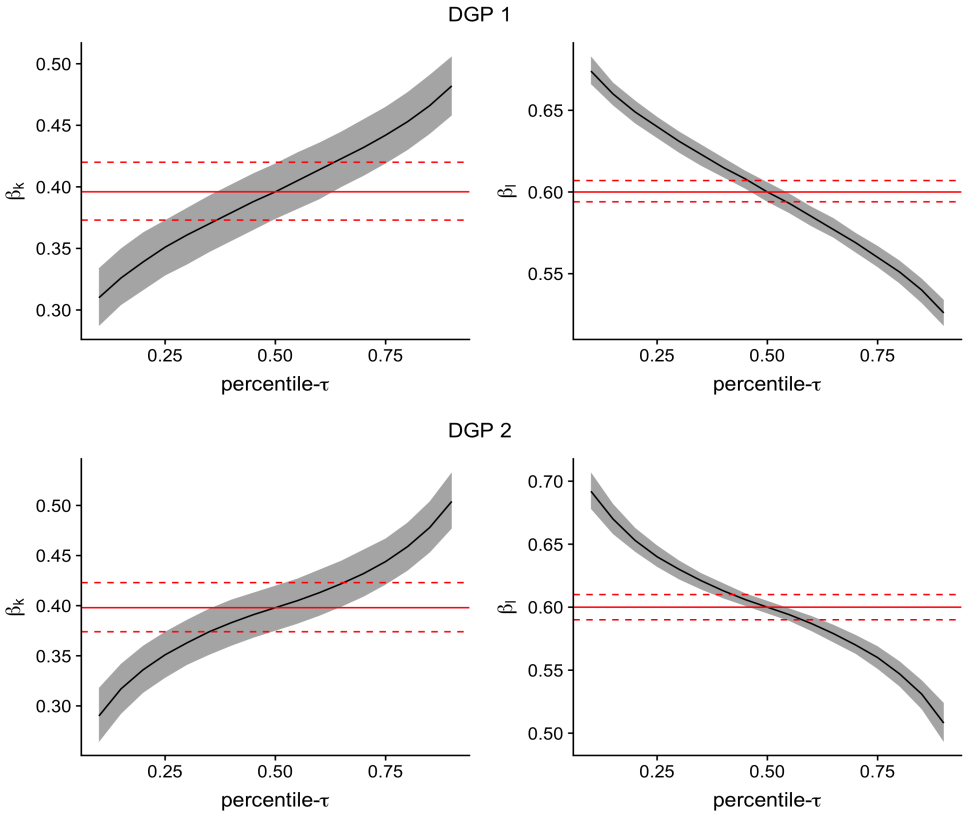
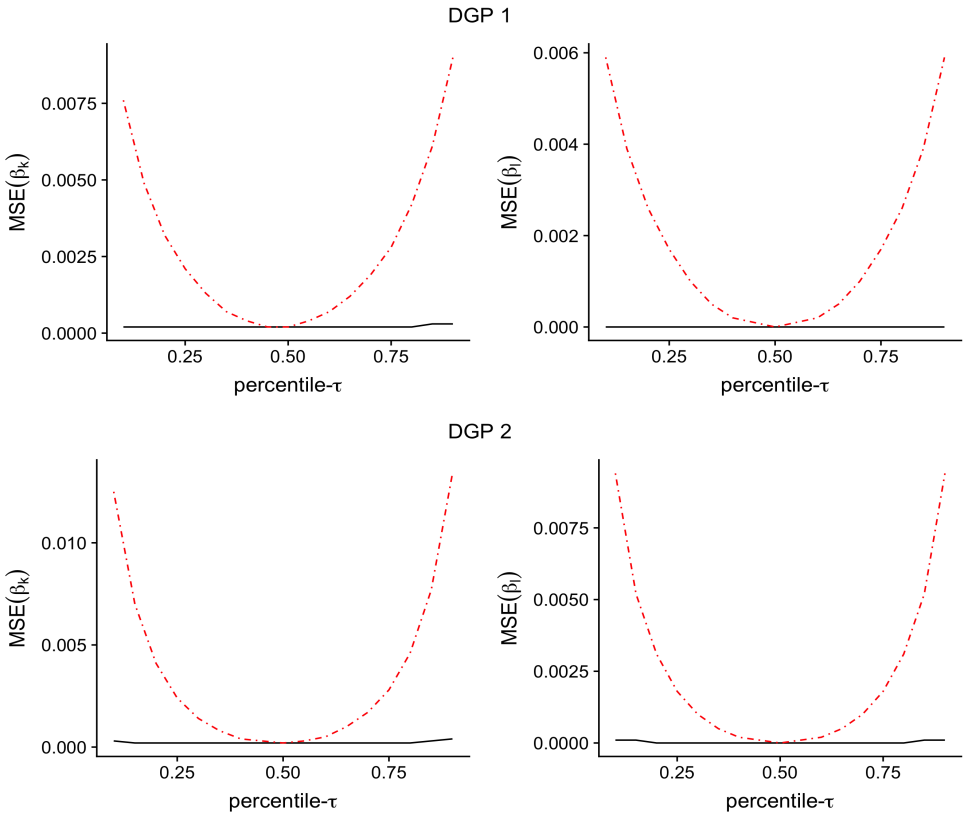


Figure 2: Simulated precision of QLP estimators of $\beta_k(\tau)$ and $\beta_l(\tau)$s. Dotted line is LP estimator.



5 Application

We apply our estimator to popular firm and plant level manufacturing datasets from the US, Chile, and Colombia to examine heterogeneity in the output distribution [8]. For each country we examine estimates across different manufacturing industries as well as how these estimates have changed over time and the implications for heterogeneity in productivity across small and large firms/plants. We use the QLP estimator presented in this paper and compare it to the LP estimates. We also compare our estimates to the quantile regression estimates without controlling for productivity. We estimate the labor coefficient using a partially regression with a 3rd degree polynomial with interactions in capital and materials. To estimate capital, we use the CUE GMM criterion function mentioned earlier with instruments $k_{it}, k_{it-1}, l_{it-1}$ and m_{it-1}. We initialize the starting values for the non-linear search using quantile regression estimates of $\beta_k(\tau)$. We use bootstrap to estimate standard errors with the number of iterations set to 500. In the first stage we use a weighted bootstrap with weights drawn from a standard exponential distribution re-sampled at each iteration. In the second stage we use a nonparametric bootstrap and recenter sample moments in each iteration.

[8] We thank Mert Demirer for providing the datasets from Chile and Colombia

5.1 US Compustat

The source for the US manufacturing data is from Compustat which covers publicly traded firms and contains data from their financial statements. We collect a sample between 1961 and 2010 on sales, capital expenditures, number of workers, and other expenses to construct measures of output, capital, labor, and material inputs using 3-digit deflators from Bartelsman and Gray (1996). Data preparation follows Keller and Yeaple (2009) and Dermirer (2020). Some issues regarding the Compustat dataset is that since the data is reported in the firm's financial statements, deflated output and input measures may not be completely capture firm's actual usage. Also, since this sample only contains publicly traded firms, it is only a fraction of all manufacturing firms in the US. Summary statistics for these deflated values are provided in Table 1. We present a series of output elasticity estimates in Table 2 which are illustrated graphically in Figures 3, 4, 5, and 6.

Table 1: Summary Statistics (in logs) for the US Manufacturing Data

Industry (NAICS code)		1st Qu.	Median	3rd Qu.	Mean	sd
31 (Total=3271)	Output	19.05	20.24	21.57	20.3	1.77
	Capital	18.66	20.37	21.76	20.19	2.12
	Labor	17.42	19.08	20.61	19.02	2.21
	Materials	17.96	19.59	21.15	19.54	2.21
32 (Total=7207)	Output	15.67	17.04	18.51	17.01	2.05
	Capital	15.65	17.51	19.13	17.31	2.41
	Labor	14.44	16.01	17.57	16.01	2.29
	Materials	14.89	16.53	18.25	16.52	2.37
33 (Total=13978)	Output	7.38	8.58	9.8	8.5	1.67
	Capital	6.67	8.29	9.74	8.15	1.95
	Labor	6.01	7.42	8.91	7.48	1.93
	Materials	6.33	7.82	9.29	7.82	1.95
All (Total=24456)	Output	18.58	19.78	21.23	19.85	1.79
	Capital	18.14	19.86	21.26	19.67	2.16
	Labor	16.98	18.59	20.13	18.56	2.17
	Materials	17.49	19.12	20.66	19.06	2.2

For each industry in the US sample, we see that the estimates of the capital elasticity is increasing in the firm size distribution and decreasing in the labor elasticity corresponding to industries 31 and 32. For industry 33 and the combined industries, the labor elasticity is an inverse U-shape; it increases quickly for low τ, but then flattens out after $\tau = 0.5$. In industries 31 and 32 only the QLP capital estimate is significantly different from the LP estimate. In industry 33 both labor and capital estimates are significantly different from the LP estimate after $\tau = 0.25$. In each industry

we compare the difference between QLP and QR estimates to test whether our model corrects for endogeneity from unobserved productivity. Bootstrap is used to construct confidence intervals of the difference between the two estimates. We find that there are significant differences between these estimates with the exception of estimates at $\tau = 0.05, 0.1$ and 0.15.

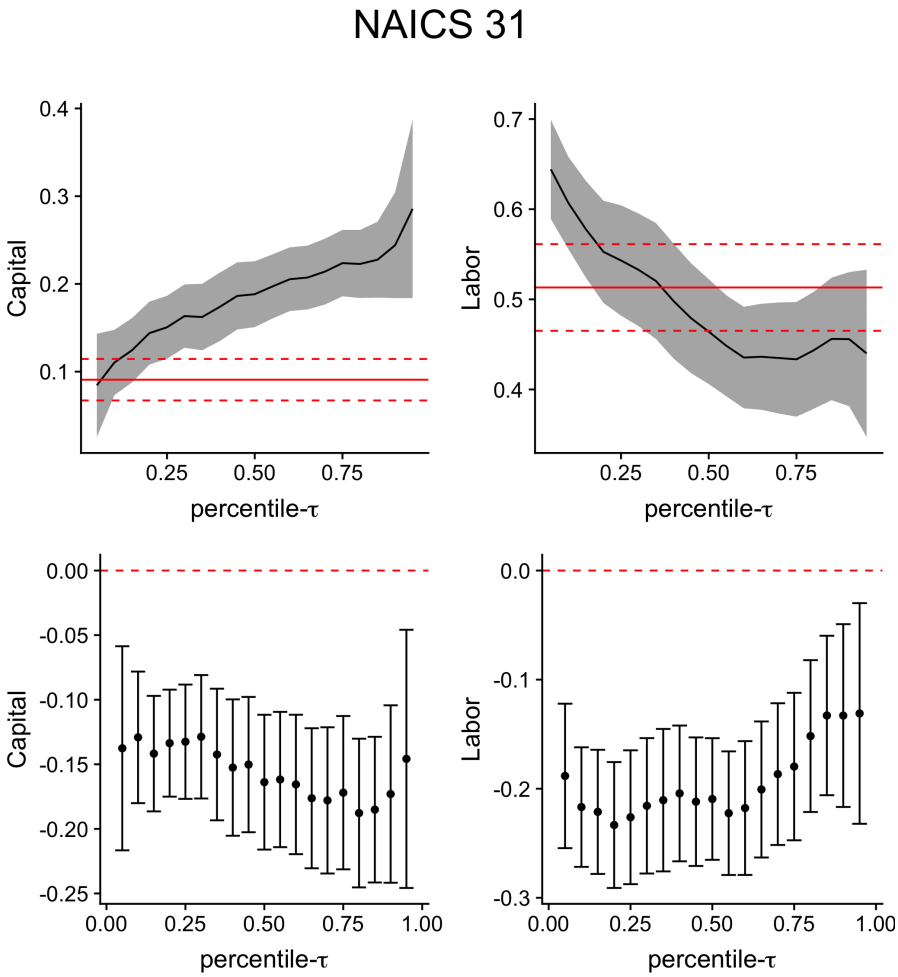


Figure 3: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

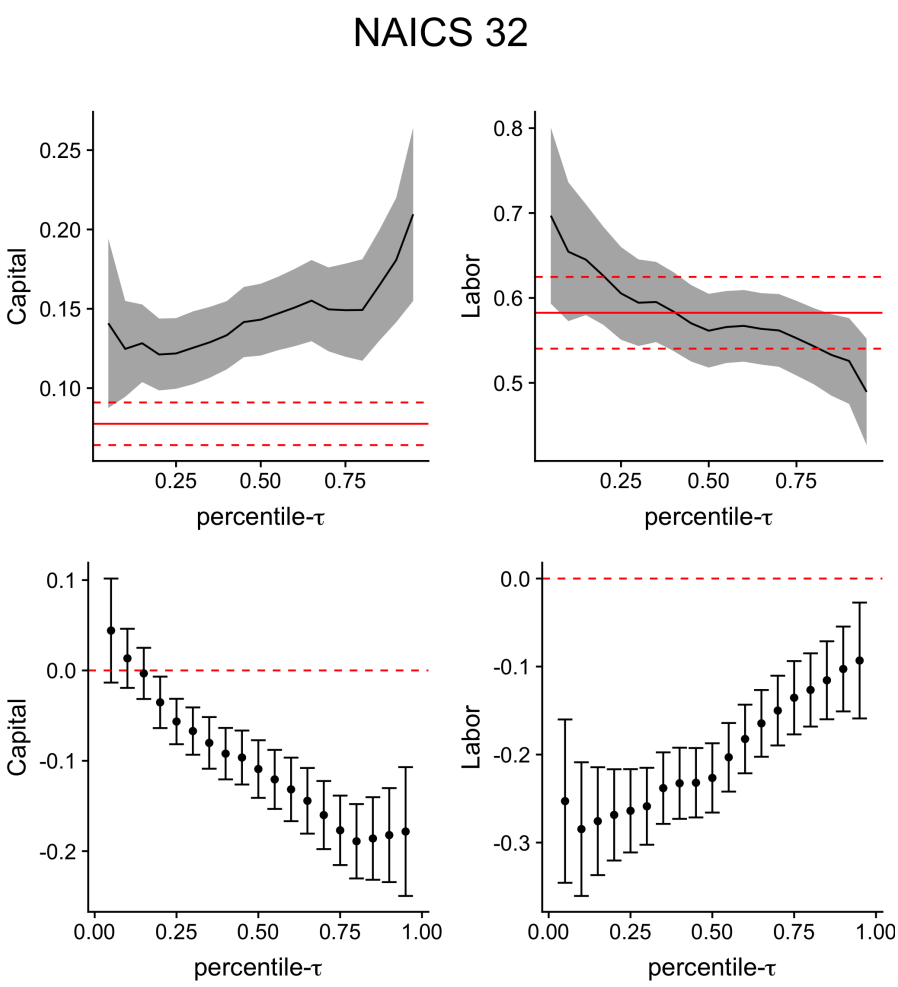


Figure 4: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

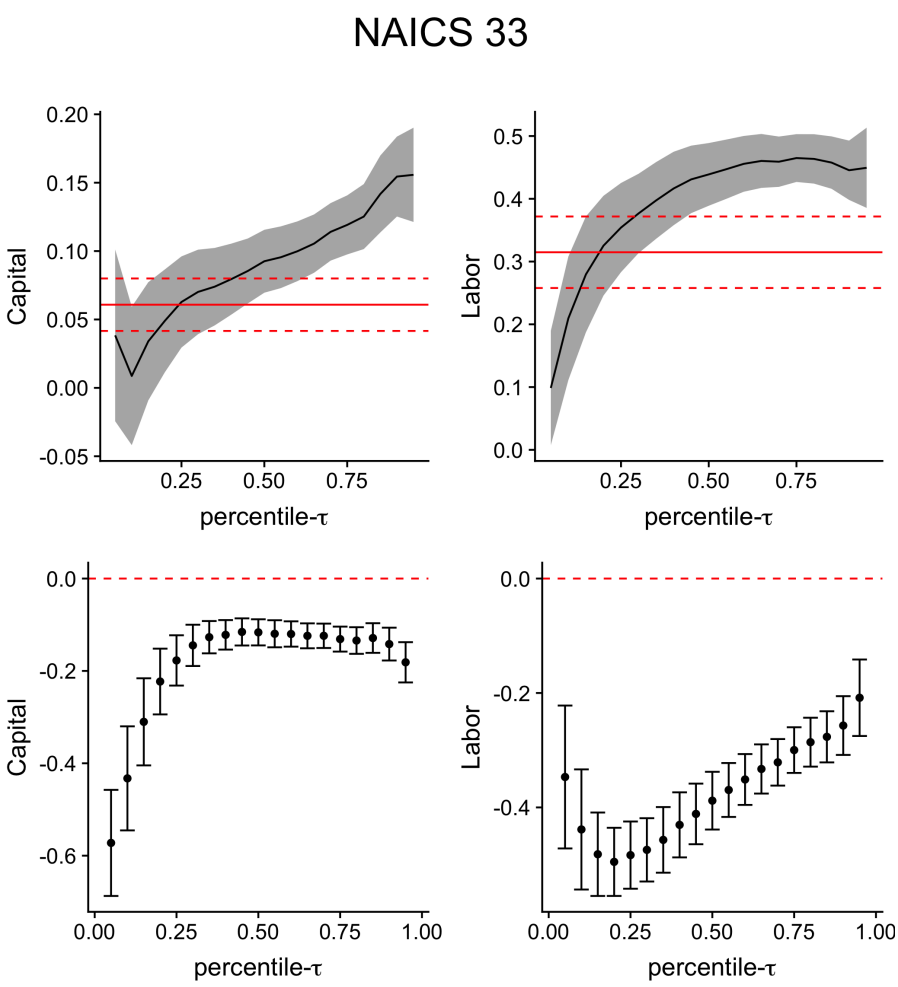


Figure 5: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

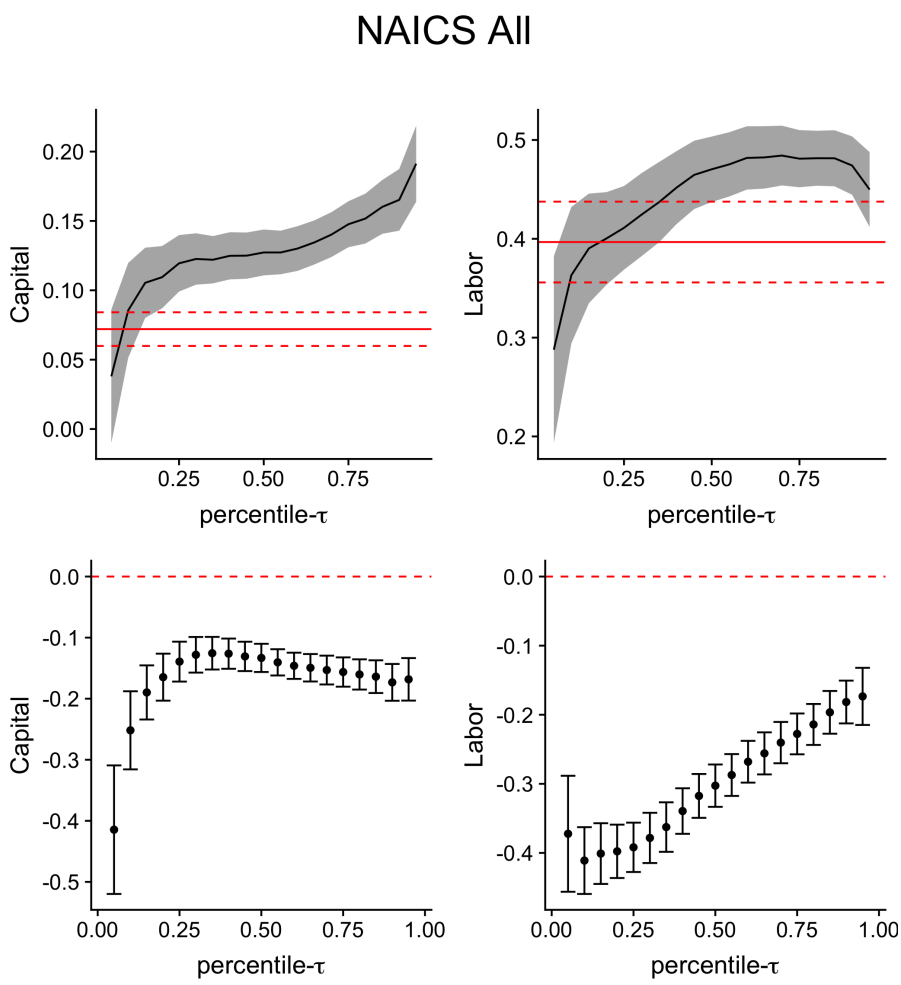


Figure 6: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

We use the estimates from the output elasticities to construct measures of returns to scale and capital intensity in Table 2. The results for returns to scale are puzzling as they are all significantly different from constant returns to scale. We can see in Industry 32 that returns to scale are generally decreasing in firm-size whereas this relationship is increasing in Industry 33. Previous papers that estimate returns to scale using the Compustat dataset such as Keller and Yeaple (2009) and Dermirer (2020) show constant returns to scale using a gross-output production function. Therefore it is possible that the empirical value-added (deflated sales minus intermediate input expenditure) is a poor proxy for value-added in our model and that value-added biases the returns to scale estimates. Differences in returns to scale in value-added and gross-output production functions are explored by Basu and Fernald (1997). We also report estimates of capital

intensity measured by the ratio of capital to labor elasticity for each quantile. In each industry, capital intensity is increasing in firm-size. This result is consistent with previous findings such as Holmes and Mitchell (2008), Kumar *et al.* (1999) and Dermirer (2020).

Table 2: Coefficient Estimates and Standard Errors for US Manufacturing Firms

		Capital		Labor		Returns to Scale		Capital Intensity	
Industry (NAICS code)	τ	Coef.	s.e.	Coef.	s.e.	Coef.	s.e	Coef.	s.e
31	0.10	0.110	0.0226	0.607	0.0311	0.717	0.0320	0.182	0.0419
	0.25	0.151	0.0218	0.543	0.0372	0.694	0.0341	0.277	0.0508
	0.50	0.188	0.0228	0.464	0.0354	0.652	0.0369	0.406	0.0633
	0.90	0.244	0.0366	0.456	0.0452	0.700	0.0489	0.535	0.1185
32	0.10	0.125	0.0184	0.654	0.0497	0.779	0.0453	0.191	0.0326
	0.25	0.122	0.0135	0.605	0.0332	0.727	0.0293	0.201	0.0277
	0.50	0.143	0.0138	0.561	0.0264	0.705	0.0269	0.255	0.0287
	0.90	0.181	0.0239	0.526	0.0307	0.707	0.0357	0.344	0.0545
33	0.10	0.009	0.0307	0.210	0.0598	0.218	0.0592	0.042	0.2655
	0.25	0.063	0.0203	0.354	0.0432	0.417	0.0429	0.177	0.0760
	0.50	0.093	0.0140	0.439	0.0304	0.531	0.0289	0.211	0.0419
	0.90	0.154	0.0178	0.446	0.0287	0.600	0.0260	0.347	0.0536
All	0.10	0.086	0.0208	0.363	0.0420	0.449	0.0505	0.236	0.0630
	0.25	0.119	0.0124	0.411	0.0257	0.531	0.0258	0.291	0.0401
	0.50	0.127	0.0100	0.470	0.0201	0.598	0.0198	0.271	0.0278
	0.90	0.165	0.0135	0.474	0.0179	0.639	0.0183	0.348	0.0361

We also use our quantile production function estimates to construct measures of firm level productivity which we define as

$$\hat{w}_{it,\tau} = \exp(y_{it} - \hat{\beta}_k(\tau)k_{it} - \hat{\beta}_l(\tau)l_{it}) \tag{42}$$

We use these measured to compare productivity growth over time to LP estimates as well as an exercise to see if there is significant dispersion in the productivity distribution over the distribution of firm size. Figure 7 reports average productivity for all US firms in the sample with the base year of the sample period set to 100. We can see that productivity growth was rapid in the beginning of the sample period but then declined after 1970 and increase again after 1980. Growth trends for each percentile of firm-size were similar although larger firms in this sample were more productive than smaller ones. Interestingly, the LP estimates are close to the productivity estimates for smaller firms at $\tau = 0.1$. This suggests that there is significant heterogeneity in the conditional firm-size distribution that conditional mean estimates of productivity such as LP cannot capture.

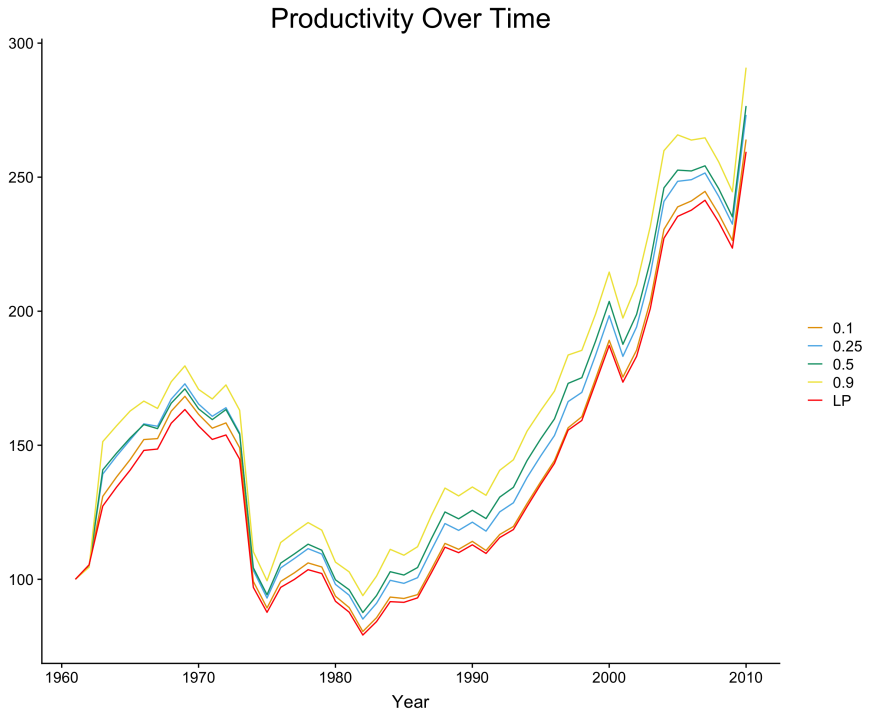


Figure 7: Estimated average productivity over time for the US. Base productivity in 1961 is set to 100.

In Figure 8 we graph estimates of certain percentiles of log-productivity distribution over the firm-size distribution for each industry in our sample. In industry 31 and 33 log-productivity is decreasing for each percentile of firm-size. Log-productivity falls faster in industry 33 compared to industry 31 for small firms, but reverses for larger firms in each industry. Log-productivity for firms in industry 32 has an inverse U-shaped relationship. It is increasing for small firms, then somewhat stagnant for medium sized firms, but then falls for very large firms. In each industry, there is substantial heterogeneity in productivity dispersion after controlling for differences in firm-size. For example, in industry 31 small firms ($\tau = 0.1$) in the 90th percentile of productivity are about 7.1 times more productive than firms in the 10th percentile of productivity. For large firms ($\tau = 0.95$), this number is about six. This suggests that there could exist some heterogeneity in common productivity dispersion estimates, such as the 90/10 ratio we use here.

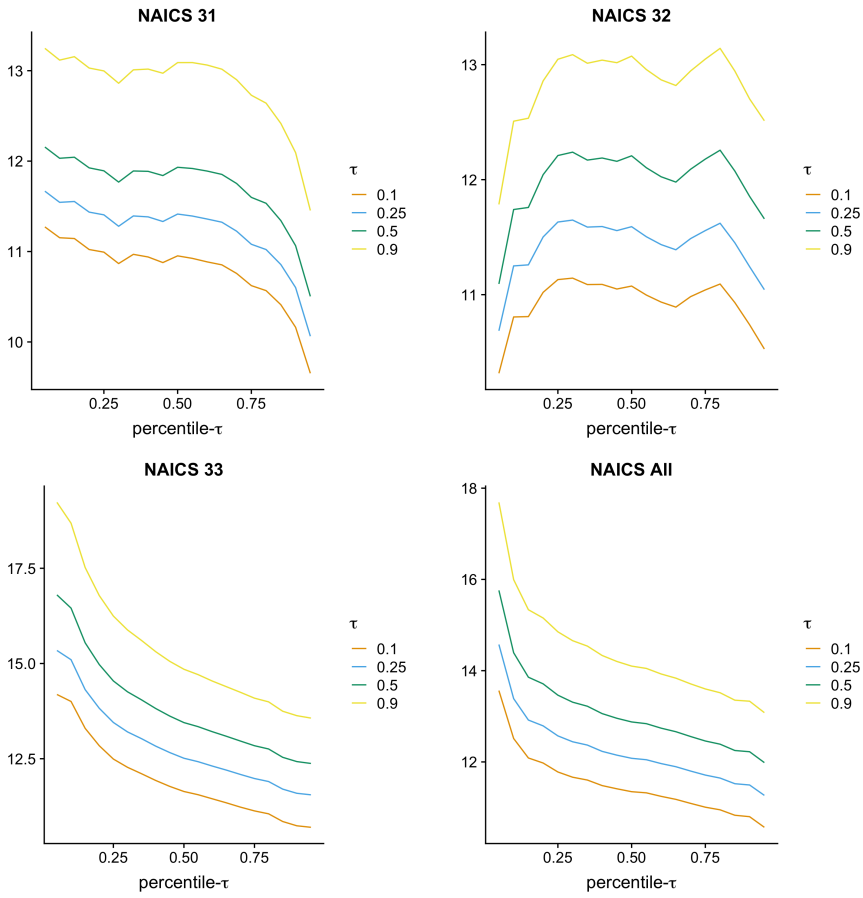


Figure 8: Estimated log-productivity at different quantiles over the firm-size distribution

We are also interested in examining the firm-size distribution over time and whether there are differences in within-firm and across firm technology heterogeneity. Figure 9 plots estimates of the output elasticities over 5 year intervals. For labor elasticity, there is heterogeneity across the firm size distribution in the beginning of the sample period. Larger firms had greater estimates of labor elasticity than very small firms. This heterogeneity decreases up until 1980 when the relationship between firm size and labor elasticity reverses. At the end of our sample period, very large firms have smaller estimates of labor elasticity than very small firms. The estimates of capital elasticity appear to be increases, however there is no discernible relationship of these estimates across different firm sizes.

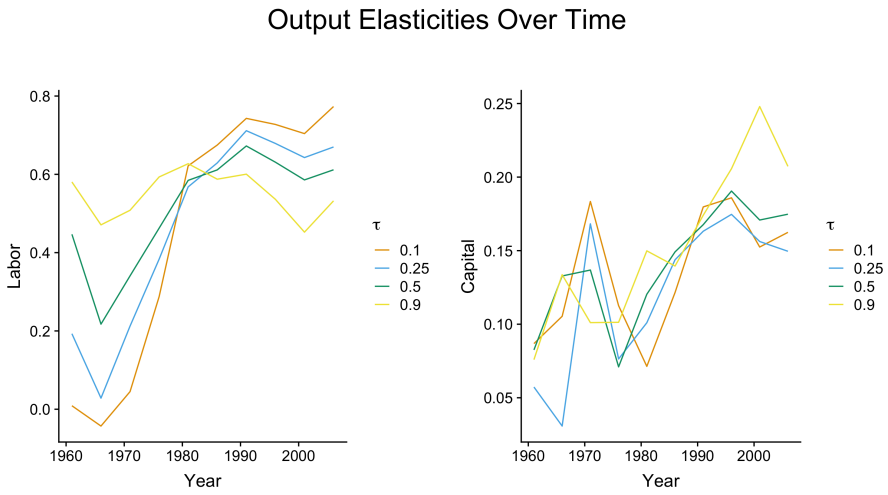


Figure 9: Estimated values of production function coefficients over time estimated at 5 year intervals

5.2 Chilean Manufacturing

This data comes from the census of Chilean manufacturing plants conducted by the Instituto Nacional de Estadistíca (INE). The sample is collected between 1979 and 1996 for firms with more than 10 employees. We divide our estimates into the three largest manufacturing industries: Food (ISIC 311), Fabricated Metals (ISIC 381), and Textiles (ISIC 321). We also aggregate the three industries with the other smaller industries to obtain estimates from the entire sample. Summary statistics for the data we use are provided in Table 3.

Figures 10, 11, and 12 illustrate our estimates from our model compared to LP estimates as well as their differences to QR estimates. Aside from ISIC 311, the estimates for labor elasticity are decreasing, but not significantly different from the LP estimates. However, since these estimates are significantly different from the QR estimate suggests that our estimator corrects for the endogeneity bias from unobserved productivity. The estimates for capital elasticity are decreasing in ISIC 311 and ISIC 381, but mostly flat for ISIC 321 and all of the industries in the sample. In every industry except for ISIC 321 there are differences between the QLP and LP estimates as well as differences between the QR estimates. These results may suggest that the productivity contributes more to heterogeneity in the estimates than do differences in the rank of the ex-post shock. From Table 4, our estimates of returns to scale are more reasonable compared to the US estimates. Interestingly returns to scale decrease as firm-size increases. Capital intensity decreases in firm-size in ISIC 311 and increases in firm-size in the entire sample.

Table 3: Summary Statistics (in logs) for Chile Manufacturing Data

Industry (ISIC code)		1st Qu.	Median	3rd Qu.	Mean	sd
311 (Total=13838)	Output	10.21	10.84	12.22	11.36	1.58
	Capital	10.56	11.4	12.4	11.52	1.37
	Labor	10.49	11.4	12.54	11.53	1.43
	Materials	10.38	11.28	12.53	11.56	1.6
381 (Total=4311)	Output	6.69	7.66	9.06	8.02	1.98
	Capital	7.52	8.51	9.7	8.65	1.68
	Labor	7.21	8.34	9.56	8.4	1.72
	Materials	7.22	8.35	9.72	8.54	1.92
321 (Total=4302)	Output	2.77	3.22	3.91	3.49	0.99
	Capital	2.89	3.47	4.22	3.71	1.08
	Labor	2.94	3.48	4.37	3.69	0.95
	Materials	2.89	3.43	4.28	3.67	1.02
All (Total=51567)	Output	9.84	10.46	11.81	10.94	1.56
	Capital	9.91	10.75	11.79	10.86	1.41
	Labor	9.68	10.62	11.75	10.73	1.48
	Materials	9.81	10.68	11.89	10.93	1.62

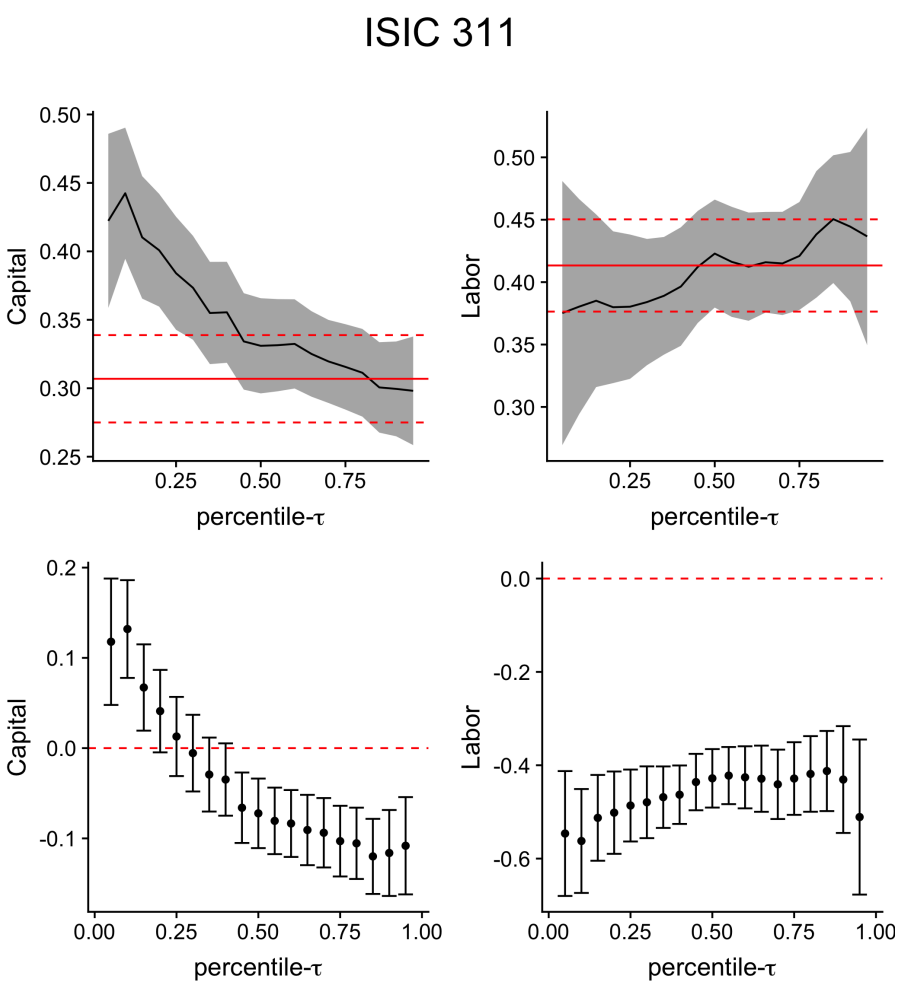


Figure 10: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

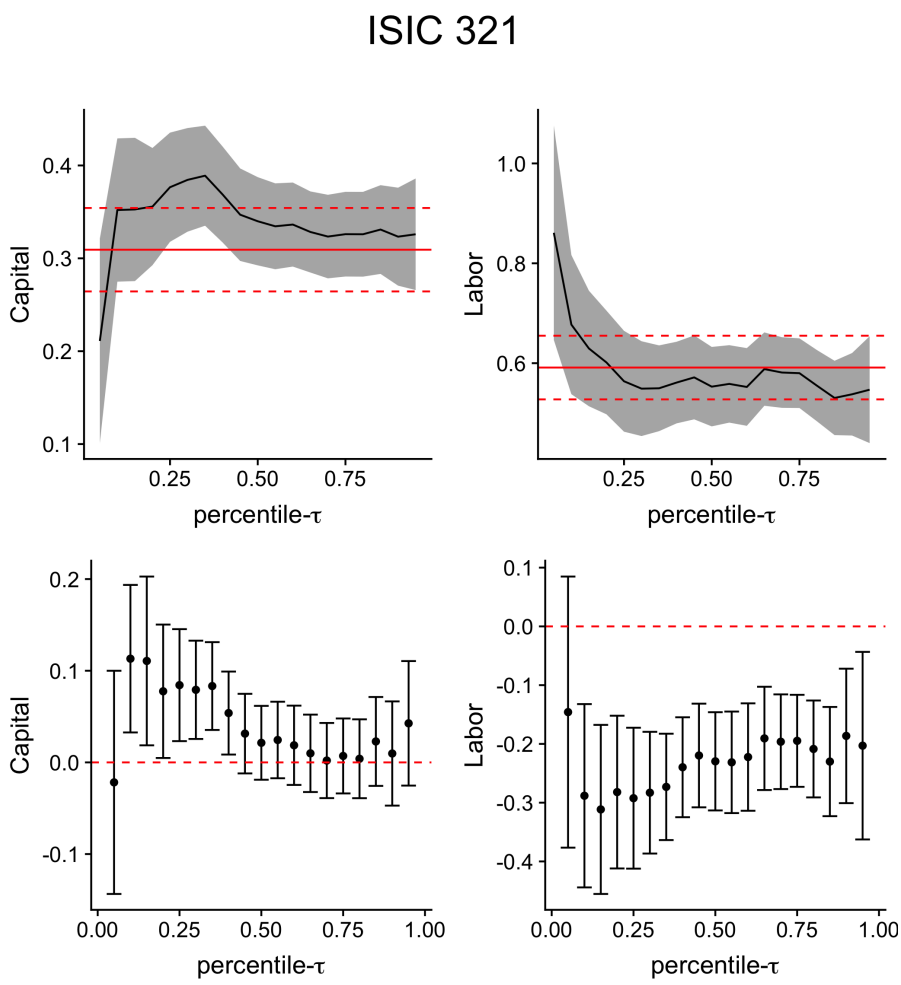


Figure 11: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

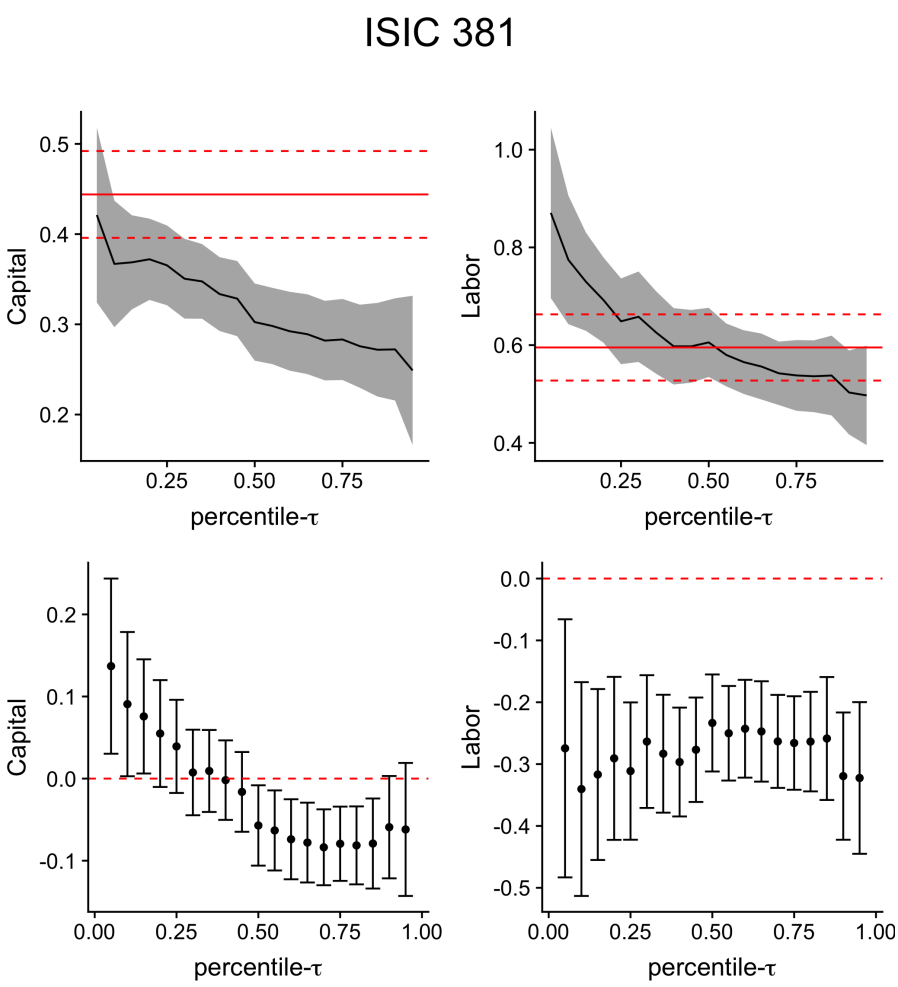


Figure 12: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

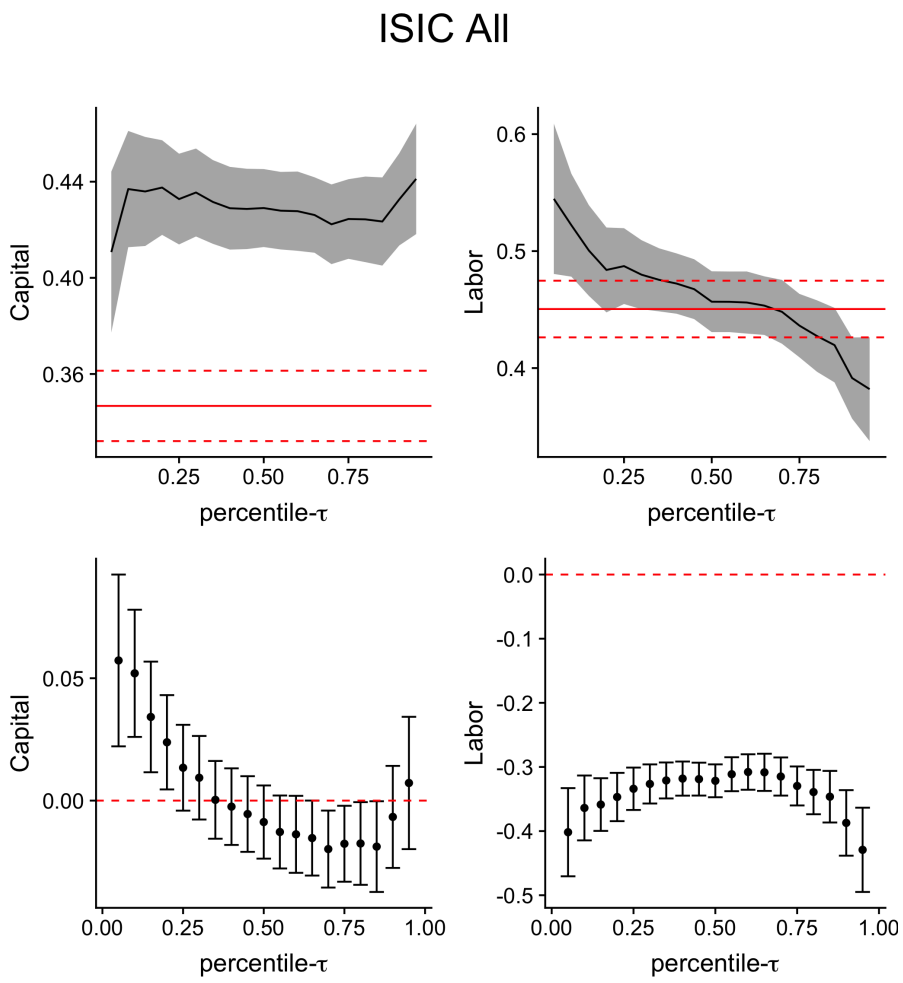


Figure 13: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

Figure 14 reports average productivity for all Chilean plants in the sample with base period set to 100. Productivity decreases in the beginning of the 1980s but the increases for the rest of the sample period. The LP estimates show higher productivity than the larger firms at $\tau = 0.9$ Figure 15 graphs the percentiles of log-productivity over the firm-size distribution for each large industry in Chile. In each industry, log-productivity is increasing in firm-size. Like the estimates in the US, there is heterogeneity in productivity dispersion. For example, in ISIC 381, small firms ($\tau = 0.1$) in the 90th percentile of productivity are about more 10.8 times productive than firms in the 10th percentile of productivity. For large firms ($\tau = 0.95$), this number is about 14.4. Lastly, Figure 16 shows the time trends in output elasticities. The estimate of labor elasticity are high for each quantile of firm size and decreases steadily with the exception of small firms ($\tau = 0.1$). The

Table 4: Coefficient Estimates and Standard Errors for Chilean Manufacturing Firms

		Capital		Labor		Returns to Scale		Capital Intensity	
Industry (ISIC code)	τ	Coef.	s.e.	Coef.	s.e.	Coef.	s.e	Coef.	s.e
311	0.10	0.442	0.0291	0.380	0.0523	0.823	0.0443	1.163	0.2123
	0.25	0.384	0.0251	0.380	0.0351	0.764	0.0307	1.010	0.1243
	0.50	0.331	0.0211	0.423	0.0263	0.754	0.0267	0.783	0.0845
	0.90	0.300	0.0211	0.444	0.0364	0.744	0.0339	0.674	0.0877
381	0.10	0.367	0.0426	0.774	0.0800	1.141	0.0594	0.474	0.1089
	0.25	0.365	0.0269	0.649	0.0533	1.014	0.0463	0.563	0.0816
	0.50	0.302	0.0260	0.606	0.0431	0.908	0.0384	0.499	0.0813
	0.90	0.272	0.0344	0.503	0.0523	0.775	0.0448	0.541	0.1273
321	0.10	0.352	0.0469	0.677	0.0847	1.029	0.0588	0.520	0.1253
	0.25	0.377	0.0357	0.564	0.0614	0.940	0.0461	0.668	0.1114
	0.50	0.340	0.0289	0.553	0.0484	0.893	0.0369	0.615	0.0961
	0.90	0.323	0.0321	0.538	0.0502	0.861	0.0412	0.601	0.1121
All	0.10	0.437	0.0147	0.522	0.0267	0.959	0.0197	0.837	0.0684
	0.25	0.433	0.0115	0.487	0.0197	0.920	0.0159	0.888	0.0547
	0.50	0.429	0.0099	0.457	0.0158	0.886	0.0134	0.939	0.0484
	0.90	0.433	0.0116	0.391	0.0211	0.824	0.0184	1.105	0.0790

estimates for capital elasticities are large aside from $\tau = 0.25$, but converge quickly after 1979.

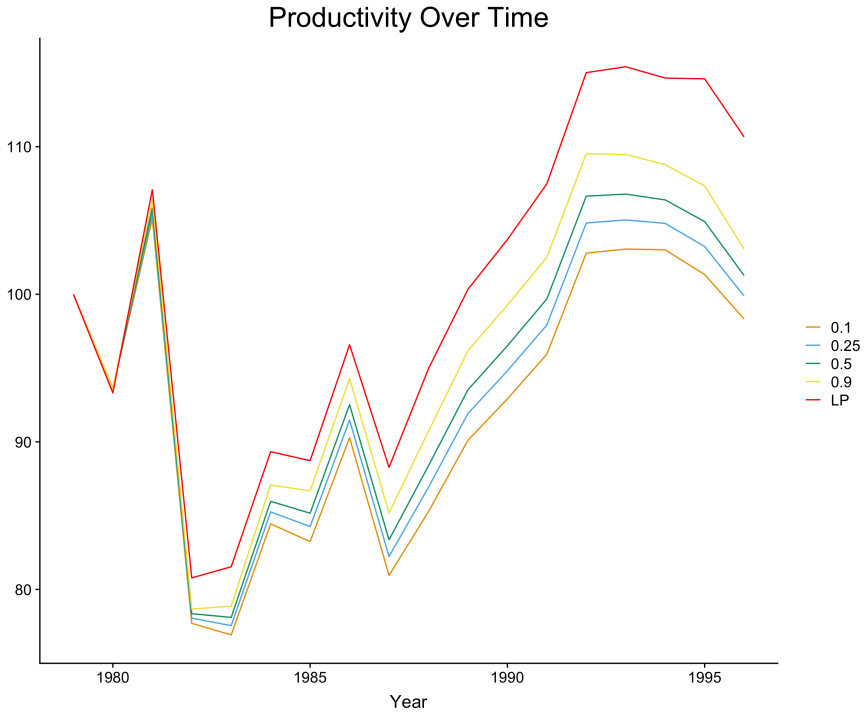


Figure 14: Estimated average productivity over time for Chile. Base productivity in 1979 is set to 100.

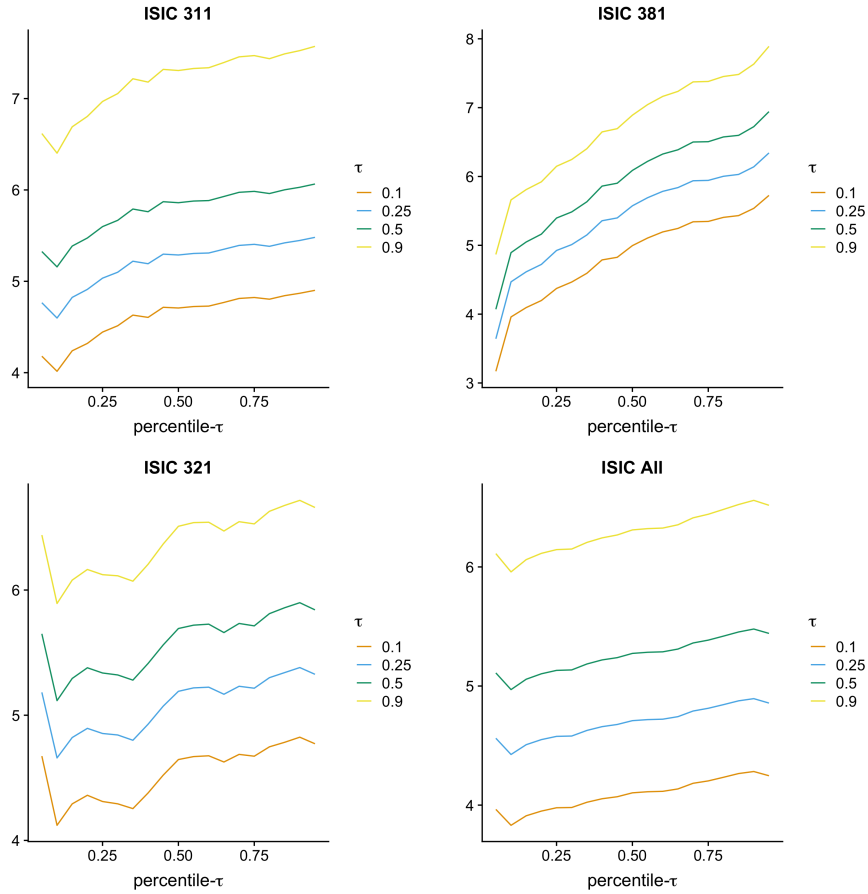


Figure 15: Estimated log-productivity at different quantiles over the firm-size distribution

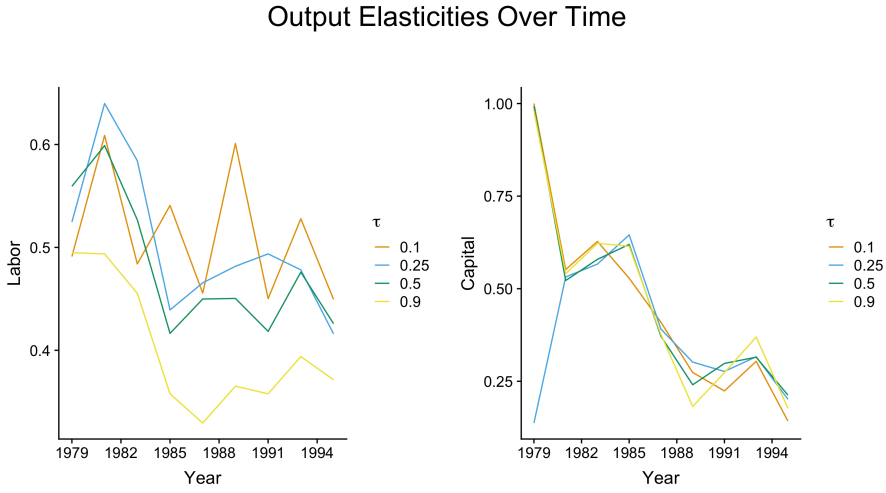


Figure 16: Estimated values of production function coefficients over time estimated at 2 year intervals

5.3 Colombian Manufacturing

This data comes is from the Colombian manufacturing census conducted by the Departamento Administrativo Nacional de Estadistica. The sample is collected between 1977 and 1991 for firms with more than 10 employees. We divide our estimates into the three largest manufacturing industries: Food (ISIC 311), Apparel (ISIC 322), and Fabricated Metals (ISIC 381). As we did with the Chilean sample, we also aggregate the three industries with other smaller industries to obtain estimates from the entire sample of manufacturing plants. Summary statistics for this data is provided in Table 5.

Table 5: Summary Statistics (in logs) for Colombia Manufacturing Data

Industry (ISIC code)		1st Qu.	Median	3rd Qu.	Mean	sd
311 (Total=13215)	Output	9.03	10.21	11.59	10.42	1.8
	Capital	8.69	9.37	10.22	9.49	1.18
	Labor	8.52	9.3	10.33	9.54	1.43
	Materials	8.7	9.62	10.88	9.92	1.67
322 (Total=12182)	Output	6.02	7.07	8.35	7.24	1.78
	Capital	5.47	6.14	6.93	6.23	1.21
	Labor	5.89	6.75	7.81	6.93	1.55
	Materials	5.9	6.89	8.16	7.12	1.77
381 (Total=7411)	Output	2.56	3.09	3.97	3.36	1.1
	Capital	2.77	3.3	3.95	3.42	0.92
	Labor	2.64	3.18	3.91	3.37	0.98
	Materials	2.71	3.3	4.11	3.5	1.09
All (Total=87783)	Output	8.39	9.73	11.26	9.87	2
	Capital	7.62	8.53	9.46	8.48	1.51
	Labor	7.77	8.65	9.72	8.8	1.58
	Materials	7.89	8.93	10.26	9.15	1.88

Figures 17, 18, and 19 illustrate estimates from our model compared to the LP estimates as well as their differences from QR estimates. The first industry, ISIC 311, shows QLP estimates of both capital and labor elasticities significantly different from LP estimates. There are also differences between these estimates and QR estimates which suggests that this method shows heterogeneity in both firm size and productivity in this industry. In ISIC 322, only the QLP estimate of labor elasticity shows significant difference from the LP estimate as well as differences from the QR estimates. There is not much difference in the QLP estimates of capital when compared to LP and QR estimates. Similar results are also true for ISIC 381. With the industries combined, both QLP estimates of capital and labor are significantly different from LP and QR estimates. For each industry, these estimates show a common trend. Capital estimates tend to increase in firm-size

whereas labor estimates tend to decrease.

Using these estimates we construct measures of returns to scale and capital intensity for each industry in Table 6. Most firms experience constant returns to scale or slightly decreasing returns to scale and we observe that returns to scale are decreasing in firm-size. The only noticeable trends in capital intensity appear in ISIC 311 and the combined sample which show decreasing relationship in firm-size in ISIC 311 and an increasing relationship in the combined sample.

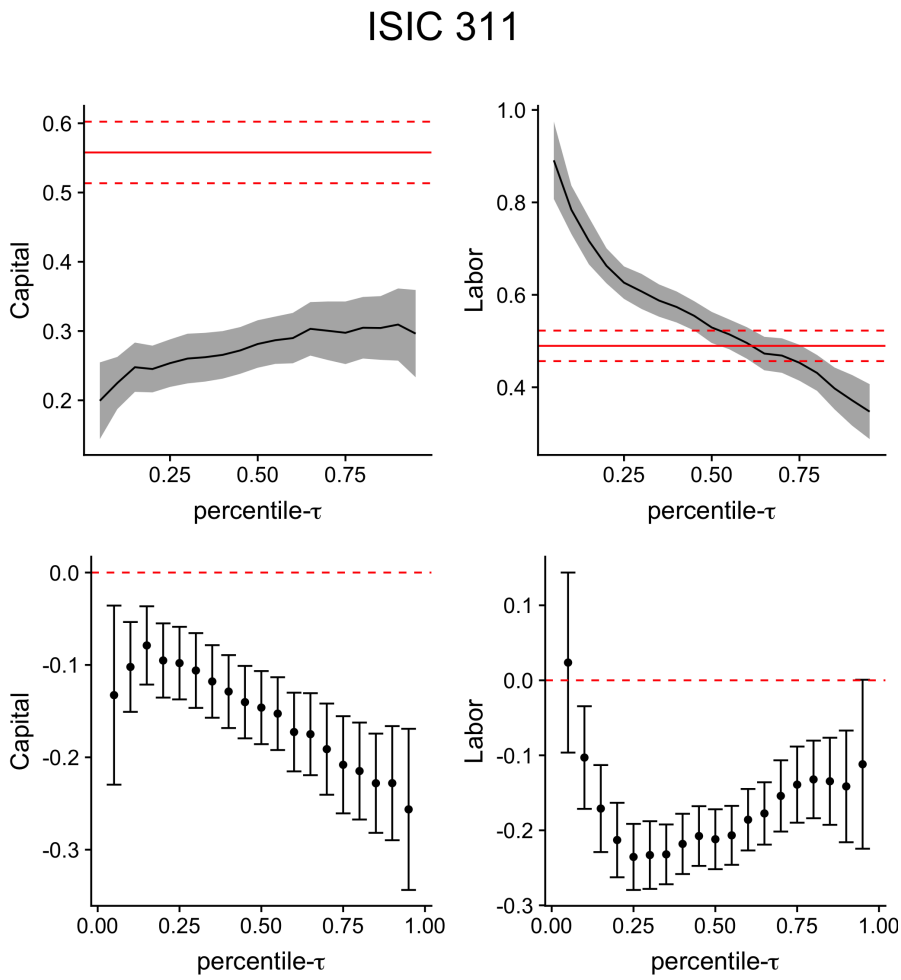


Figure 17: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

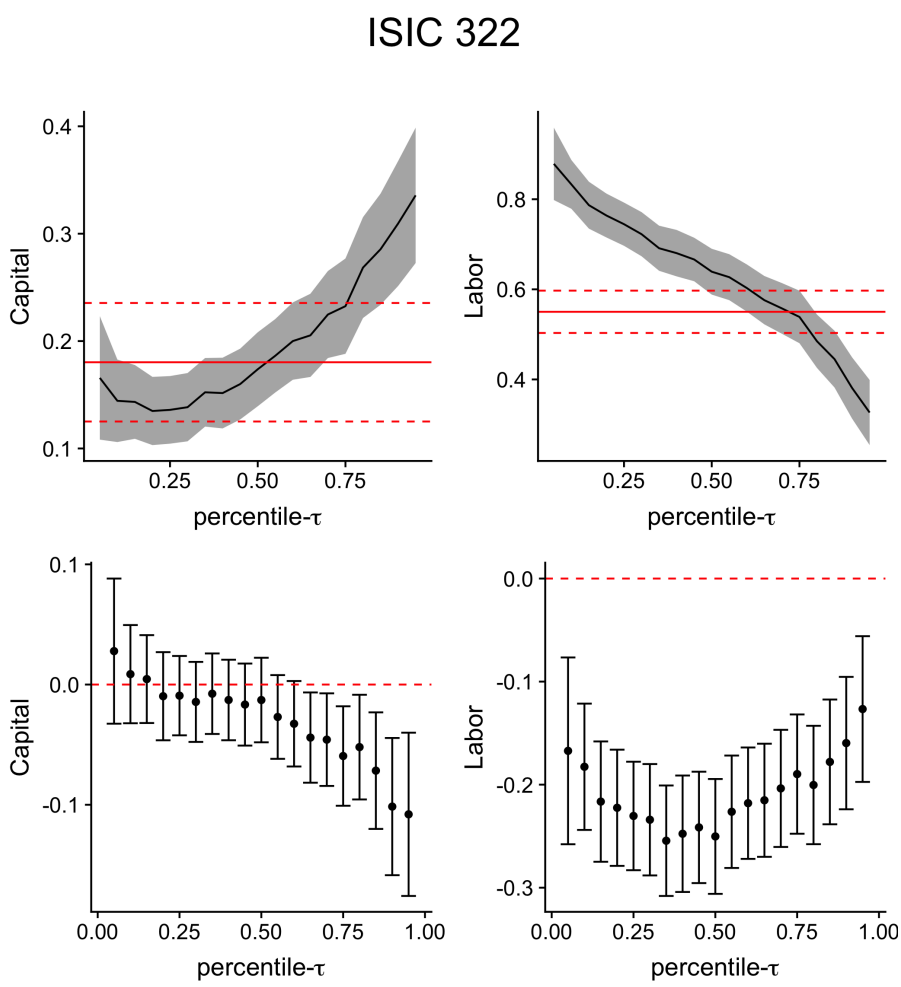


Figure 18: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

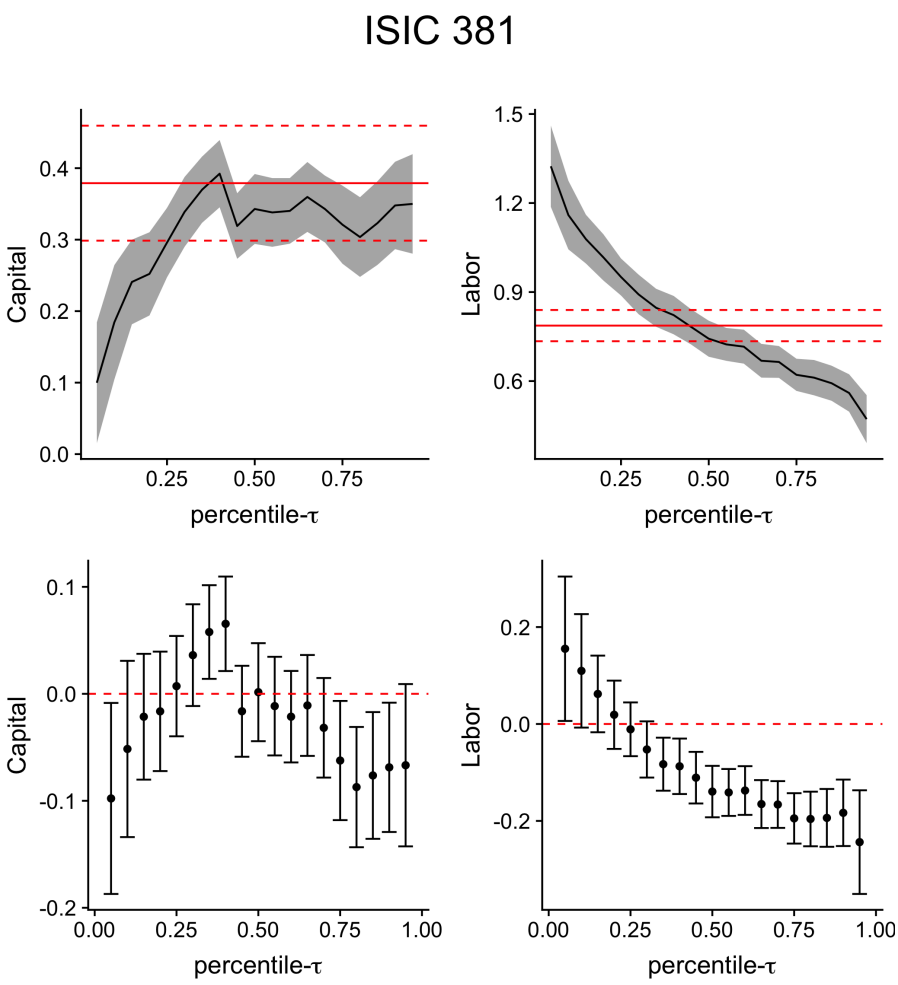


Figure 19: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

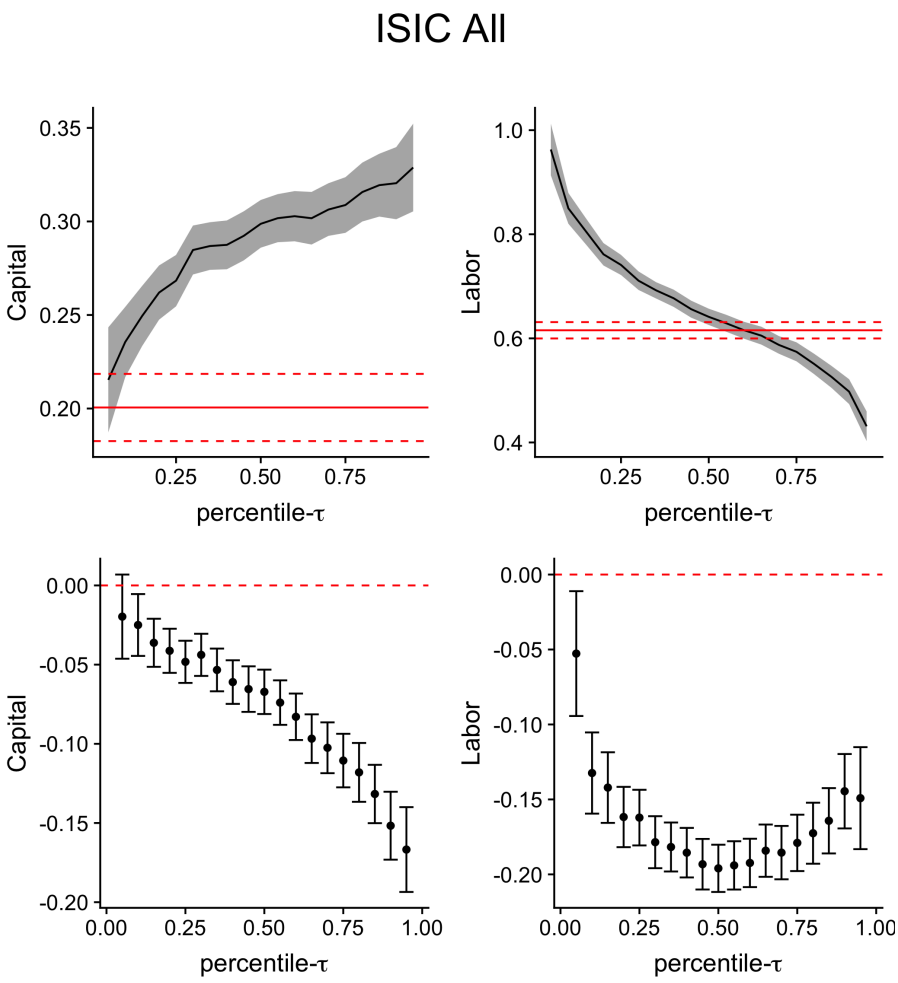


Figure 20: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

Figure 21 reports average productivity for all Colombian plants in the sample with base period set to 100. Productivity decreases in the beginning of the sample period but then increases for the rest of the sample period after 1980 with some sharp periods of productivity decline and incline. Each percentile of firm size has similar productivity levels at the beginning of the sample period, but diverge after 1984. The LP estimates show just higher than productivity of small firms at $\tau = 0.1$ Figure 22 graphs the percentiles of log-productivity over the firm-size distribution for each large industry in Colombia. In each industry, log-productivity is increasing in firm-size. Like the estimates in the US and Chile, there is heterogeneity in productivity dispersion. For example, in ISIC 322, small firms ($\tau = 0.1$) in the 90th percentile of productivity are about 4.9 times more productive than firms in the 10th percentile of productivity. For large firms ($\tau = 0.95$), this number

Table 6: Coefficient Estimates and Standard Errors for Colombian Manufacturing Firms

		Capital		Labor		Returns to Scale		Capital Intensity	
Industry (ISIC code)	τ	Coef.	s.e.	Coef.	s.e.	Coef.	s.e	Coef.	s.e
311	0.10	0.225	0.0229	0.784	0.0318	1.009	0.0274	0.287	0.0384
	0.25	0.254	0.0209	0.626	0.0214	0.880	0.0222	0.405	0.0422
	0.50	0.281	0.0208	0.529	0.0205	0.811	0.0223	0.531	0.0546
	0.90	0.309	0.0317	0.372	0.0333	0.681	0.0348	0.832	0.1648
322	0.10	0.144	0.0233	0.833	0.0329	0.977	0.0248	0.173	0.0338
	0.25	0.136	0.0191	0.745	0.0292	0.881	0.0225	0.183	0.0319
	0.50	0.174	0.0209	0.639	0.0309	0.813	0.0215	0.272	0.0460
	0.90	0.309	0.0354	0.381	0.0413	0.690	0.0311	0.812	0.1728
381	0.10	0.184	0.0486	1.159	0.0703	1.344	0.0400	0.159	0.0507
	0.25	0.296	0.0295	0.951	0.0379	1.247	0.0268	0.311	0.0428
	0.50	0.343	0.0298	0.743	0.0369	1.086	0.0356	0.462	0.0541
	0.90	0.348	0.0372	0.560	0.0383	0.908	0.0403	0.621	0.1010
All	0.10	0.236	0.0114	0.850	0.0179	1.085	0.0154	0.277	0.0165
	0.25	0.268	0.0083	0.741	0.0118	1.009	0.0095	0.362	0.0157
	0.50	0.299	0.0077	0.641	0.0095	0.940	0.0089	0.466	0.0169
	0.90	0.320	0.0117	0.498	0.0145	0.818	0.0136	0.644	0.0402

is about 6.7. Finally, Figure 23 shows the time trends in output elasticities. The estimate of labor elasticity are about 0.6 for each quantile of firm size and increases steadily until about 1981 then starts to decrease. At the end of the sample period there is more heterogeneity in these estimates. Capital elasticity estimates tend to decrease during the sample period for each quantile of firm size.

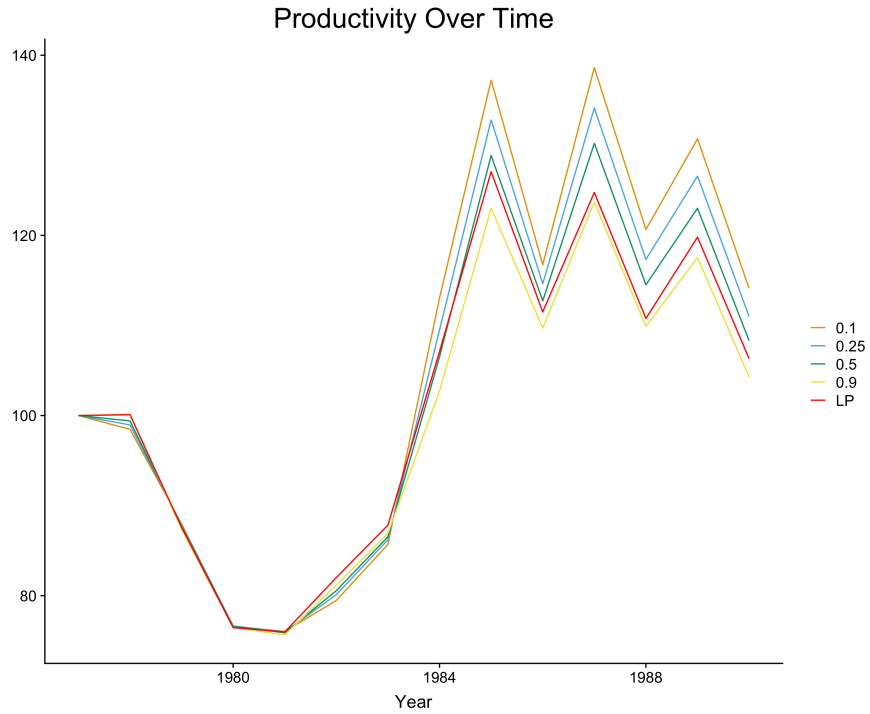


Figure 21: Estimated average productivity over time for Colombia. Base productivity in 1978 is set to 100.

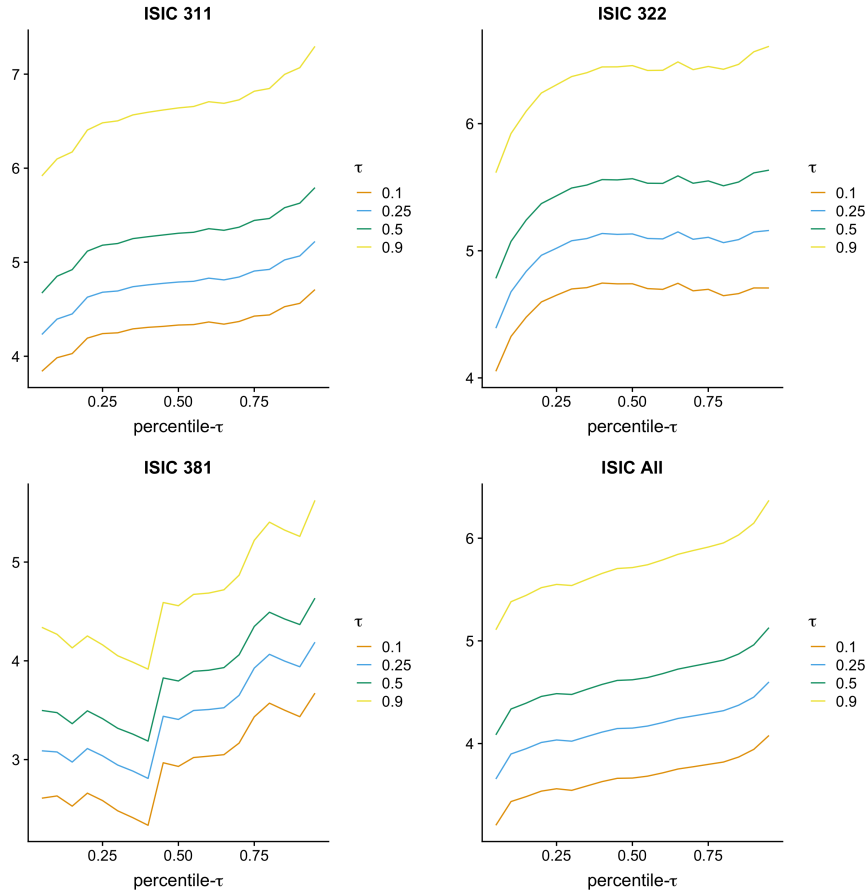


Figure 22: Estimated log-productivity at different quantiles over the firm-size distribution

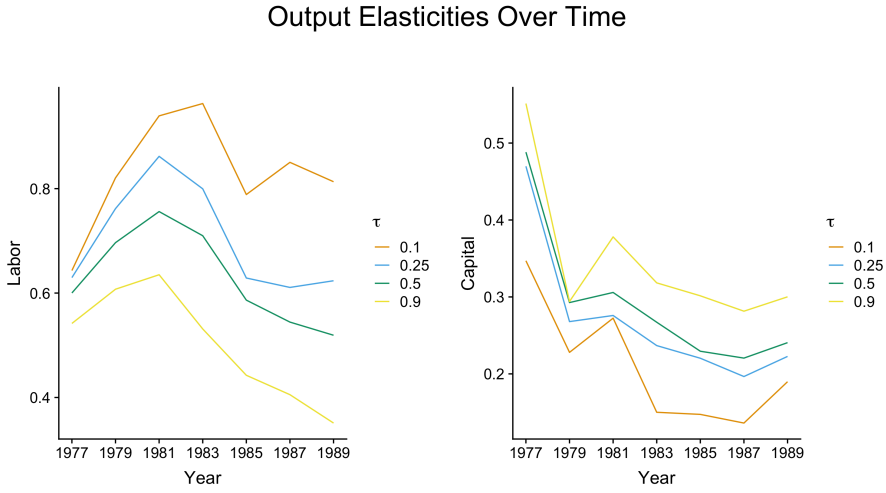


Figure 23: Estimated values of production function coefficients over time estimated at 2 year intervals

6 Conclusions

We proposed a method that extends the intermediate input proxy variable approach to estimating quantiles of the conditional quantiles of firm production. The method is computationally attractive as it resembles the two-stage estimator introduced in the control function literature with conditional quantile restrictions in the first stage. As a result, practitioners are able to easily apply the proposed estimator to production function models where the data reveal significant heterogeneous output elasticities along the conditional firm-size distribution. We showed that identification conditions for second stage estimates based on conditional quantiles are not as straightforward as conditional mean moment restrictions in the presence of two unobservables. Using the concentrated moment approach of Ackerberg *et al.* (2015) appears to fix this issue. We showed that this estimator works well in finite samples and showed that it captures heterogeneity in firm-size under different data generating processes. An application to widely used datasets from the US, Chile, and Colombia showed that in some industries, our estimator captures unobserved heterogeneity that the LP estimator does not.

Improvements and extensions of this estimator are currently being explored. For example, using a value-added production function may show more heterogeneity in estimates of elasticities and productivity than a gross-output production function. However, using a gross-output production function with an intermediate input proxy variable suffers from non-identification problems. Therefore, a structural value-added production function may be preferable. Kasahara *et al.* (2017) show how to modify OP/LP type moment conditions for a value-added production function from a gross-output production function. It would be interesting to explore whether those methods could be applied here. This paper also makes an interesting connection between the literature on production risk and quantile utility maximization. Currently, quantile utility maximization problems and estimation of these models are being studied by de Castro and Galvao (2017) and de Castro *et al.* (2018) in the context of dynamic consumption problems. It would be interesting to explore a model for a firm who maximizes quantile utility of profits which could provide an alternative explanation for unobserved heterogeneity from quantile regression estimates.

This paper contributes to the growing literature on production functions with unobserved heterogeneity. We show that differences in firm-size correspond to the the rank of the ex-post shock. The control function approach used here restricts us from examining other dimensions of firm heterogeneity. For example, this approach only allows us to use a location shift model for productivity. The consequence of this is that estimated growth rates in productivity are the same across firm-size which may not be true in practice. Allowing richer distributional affects of productivity would be an interesting extension. This approach also restricts us from examining non-Hicks neutral productivity shocks factor-augmenting productivity. We are currently working on an extension of this paper to a non-separable model to address these last two points, but the estimator we propose here is computationally attractive and easy to implement in empirical research.

References

Abrevaya, J. and Dahl, C. M. (2008). The effects of birth inputs on birthweight. *Journal of Business & Economic Statistics*, **26** (4), 379–397.

Ackerberg, D., Caves, K. and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, **83** (6), 2411–2451.

—, Chen, X. and Hahn, J. (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics*, **94** (2), 481–498.

—, —, — and Liao, Z. (2014). Asymptotic efficiency of semiparametric two-step GMM. *The Review of Economic Studies*, **81** (3), 919–943.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, **71** (6), 1795–1843.

— and — (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, **170** (2), 442–457.

Aigner, D., Lovell, C. and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, **6** (1), 21–37.

Antle, J. M. (1983). Testing the stochastic structure of production: A flexible moment-based approach. *Journal of Business & Economic Statistics*, **1** (3), 192–201.

Aragon, Y., Daouia, A. and Thomas-Agnan, C. (2005). NONPARAMETRIC FRONTIER ESTIMATION: A CONDITIONAL QUANTILE-BASED APPROACH. *Econometric Theory*, **21** (02).

Bache, S. H. M., Dahl, C. M. and Kristensen, J. T. (2012). Headlights on tobacco road to low birthweight outcomes. *Empirical Economics*, **44** (3), 1593–1633.

Balat, J., Brambilla, I. and Sasaki, Y. (2018). Heterogeneous firms: skilled-labor productivity and the destination of exports, Working paper.

Bartelsman and Gray (1996). *The NBER Manufacturing Productivity Database*. Tech. Rep. 205, National Bureau of Economic Research.

Basu, S. and Fernald, J. G. (1997). Returns to scale in u.s. production: Estimates and implications. *Journal of Political Economy*, **105** (2), 249–283.

Bernini, C., Freo, M. and Gardini, A. (2004). Quantile estimation of frontier production function. *Empirical Economics*, **29** (2), 373–381.

Bhattacharya, D. (2009). Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association*, **104** (486), 486–500.

Cai, Z., Chen, L. and Fang, Y. (2018). A semiparametric quantile panel data model with an application to estimating the growth effect of FDI. *Journal of Econometrics*, **206** (2), 531–553.

Canay, I. A. (2011). A simple approach to quantile regression for panel data. *The Econometrics Journal*, **14** (3), 368–386.

Chamberlain, G. (1984). Panel data. In: Griliches Z, Intriligator MD (eds) Handbook of econometrics. vol. 2, Elsevier Science B.V., New York, pp. 1247–1318.

Chambers, C. P. (2007). Ordinal aggregation and quantiles. *Journal of Economic Theory*, **137** (1), 416–431.

Chaudhuri, P. (1991a). Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of Multivariate Analysis*, **39** (2), 246–269.

— (1991b). Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of Statistics*, **19** (2), 760–777.

Chen and Pouzo (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, **80** (1), 277–321.

Chen, X. and Pouzo, D. (2009). Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, **152** (1), 46–60.

Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, **73** (1), 245–261.

de Castro, L., Galvao, A. F., Kaplan, D. M. and Liu, X. (2018). Smoothed GMM for quantile models. *Journal of Econometrics*, forthcoming.

de Castro, L. I. and Galvao, A. F. (2017). Dynamic quantile models of rational behavior. *SSRN Electronic Journal*.

Dermirer, M. (2020). Production function estimation with factor augmenting technology: An application to markups.

Gandhi, A., Navarro, S. and Rivers, D. A. (2020). On the identification of gross output production functions. *Journal of Political Economy*, pp. 000–000.

Griliches, Z. and Hausman, J. A. (1986). Errors in variables in panel data. *Journal of Econometrics*, **31** (1), 93–118.

Hausman, J., Liu, H., Luo, Y. and Palmer, C. (2019). *Errors in the Dependent Variable of Quantile Regression Models*. Tech. rep.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, **51** (2), 186.

Holmes, T. J. and Mitchell, M. F. (2008). A theory of factor allocation and plant size. *The RAND Journal of Economics*, **39** (2), 329–351.

Just, R. E. and Pope, R. D. (1978). Stochastic specification of production functions and economic implications. *Journal of Econometrics*, **7** (1), 67–86.

— and — (1979). Production function estimation and related risk considerations. *American Journal of Agricultural Economics*, **61** (2), 276–284.

Kaplan, D. M. and Sun, Y. (2016). Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory*, **33** (1), 105–157.

Kasahara, H., Schrimpf, P. and Suzuki, M. (2017). Identification and estimation of production function with unobserved heterogeneity, Working paper.

Keller, W. and Yeaple, S. R. (2009). Multinational enterprises, international trade, and productivity growth: Firm-level evidence from the united states. *Review of Economics and Statistics*, **91** (4), 821–831.

Koekner, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, **81** (4), 673–680.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91** (1), 74–89.

Kumar, K., Rajan, R. and Zingales, L. (1999). *What Determines Firm Size?* Tech. rep.

Lamarche, C. (2010). Robust penalized quantile regression estimation for panel data. *Journal of Econometrics*, **157** (2), 396–408.

Lee, S. (2003). EFFICIENT SEMIPARAMETRIC ESTIMATION OF a PARTIALLY LINEAR QUANTILE REGRESSION MODEL. *Econometric Theory*, **19** (01).

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *Review of Economic Studies*, **70** (2), 317–341.

Li, T. and Sasaki, Y. (2017). Constructive identification of heterogeneous elasticities in the cobb-douglas production function, Working paper.

Manski, C. F. (1988). Ordinal utility models of decision making under uncertainty. *Theory and Decision*, **25** (1), 79–104.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, **64** (6), 1263.

Rosen, A. M. (2012). Set identification via quantile restrictions in short panels. *Journal of Econometrics*, **166** (1), 127–137.

Rostek, M. (2009). Quantile maximization in decision theory. *The Review of Economic Studies*, **77** (1), 339–371.

Schennach, S. (2014). Entropic latent variable integration via simulation. *Econometrica*, **82** (1), 345–385.