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

Justin Doty*and Suyong Song†

February 26, 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 both conditional mean and quantile restrictions in a two-stage approach. We show that this method allows us to capture heterogeneity in output elasticities along the firm-size distribution that may not be found in conditional mean estimates of Olley and Pakes (1996) and Levinsohn and Petrin (2003). 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) (OP) or an intermediate material input using Levinsohn and Petrin (2003) (LP) or Ackerberg *et al.* (2015) (ACF). These proxies are functions of a state variable such as capital and unobserved productivity. Under certain assumptions, this function is strictly increasing in its scalar unobserved productivity component. Inverting this function controls

*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

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 heterogeneity in firms behaviors and decisions.[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 for firm heterogeneity in production technology beyond a Hick's neutral productivity shock to be driven by the rank of the unobserved ex-post production shock. We use the proxy variable approach in this framework in order to control for the part of production unobservables that are correlated with input decisions. The literature on control function approaches for quantile regression models is relatively small so it is not straightforward to estimate production functions by allowing for endogenous inputs and their heterogeneous coefficients.[2] 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 popular control function approach of Levinsohn and Petrin (2003) and is successful in both capturing heterogeneous output elasticities and controlling for unobserved productivity. In our empirical application, we consider several popular firm and plant-level manufacturing datasets and compare our estimator to the LP estimator. We show that heterogeneity in these estimates implies differences in other features of firm production, such as returns to scale, capital intensity and productivity.

The rest of the paper is organized as follows. Section 2 reviews prior approaches for production function estimation and the interpretations of quantile production functions. 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.

[2] See for example Chesher (2003), Ma and Koenker (2006), Lee (2007) and Imbens and Newey (2009)

2 Literature Review

2.1 Production Function Estimation

We briefly review the LP (2003) procedure for estimating a *value-added* production function (in logs) [3],[4].

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

where y_{it} denotes value-added output for firm i at time t, l_{it} denotes labor input, 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 introduces an intermediate input demand defined as [5]

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

where the function m 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 this equation into the production function they obtain

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

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. First stage estimates of β_l and Φ can be obtained by a local linear regression or a polynomial regression in (k_{it}, m_{it}).

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 $\mathbb{E}[\xi_{it} | \mathcal{I}_{it-1}] = 0$.

[3] We consider a value-added production function here to be consistent with the model we introduce in Section 3

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

[5] 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

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

where $\hat{y}_{it} = y_{it} - \hat{\beta}_l l_{it}$ 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 estimator complicates asymptotic inference.

2.2 Production Functions and Quantile Regression

A quantile regression framework for production functions may facilitate many different interpretations. Since this framework has been applied in the production frontier literature, we briefly review how this may be a natural interpretation and the limitations of such a model. 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. Bernini *et al.* (2004) suggests quantile regression could be used to estimate the highest percentiles of the conditional output distribution as an estimate of the production frontier. Aragon *et al.* (2005) uses a characterization of a deterministic frontier that interprets production functions as being from a continuous interval $\tau \in [0, 1]$ where the $\tau = 1$ corresponds to the efficient frontier. One difficulty with these approaches is choosing which quantile to estimate as the frontier. Predicting differences between the frontier and a production process for a given firm is also complicated since the predicted error is a composite of technical inefficiency and noise. Other econometric issues such as endogeneity of input choices with respect to inefficiency are difficult to incorporate in this framework. The purpose of this paper is not to debate the advantages and disadvantages of production frontier models so we leave this discussion for future studies and acknowledge the theoretical challenges of quantile frontier models and instead focus on the standard production function model.

There are two main challenges with implementing a quantile regression framework to the standard production function model. First, addressing the endogeneity of ω_{it} using traditional panel

data methods or control function approaches for quantile models is still a developing field that faces many challenges such as incidental parameters or non-smooth criterion functions. Second, it is not straightforward to link firms output decisions to variations in the ex-post shock η_{it} without imposing some strict structural model of firm production.

Addressing the first point, we will argue that recent quantile panel data and quantile IV models should not be used for the same reasons panel data and IV models are not used for the conditional mean model. 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), or Canay (2011) which assumes the fixed effect shifts only the location of the conditional quantile function. This approach has two disadvantages. First, assuming the unobservable is time-invariant is restrictive. Griliches and Hausman (1986) show that a fixed effect for productivity, ω_i, leads to low estimates of the capital elasticity. Second, 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). This correlated random effect (CRE) approach was used by Abrevaya and Dahl (2008) and variations of this type of estimator have been developed by other researchers. One issue with the CRE approach is that identification of the conditional quantile function is difficult because it now 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 using

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

We do not use this procedure since input prices may not have enough variation across firms and exogeneity can be violated if they capture input quality differences as shown by Griliches and Hausman (1986).

3 A Random Coefficient Production Function

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

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

The variables in equation (10) have the same interpretation as the ones we introduced in the LP model. In this specification we allow the output elasticities to be functionally dependent on the production shock η_{it} while productivity ω_{it} still maintains its additive separability. 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 special structural production functions such as Leontief value-added. Since the production shock η_{it} enters (10) non-separably, it is difficult to recover gross-output objects from value-added since the latter is composed of the additional error term. One consequence of a value-added model is that estimates of productivity may be more disperse than gross-output estimates which controls for material inputs. Therefore, in our empirical application we interpret the elasticities and productivity with some caution.

One advantage of the value-added model is that the rank of η_{it} can be interpreted as the rank of firm-size measured by output net of material inputs. This way, we are able to measure firm-size by the value of the firm's contribution to output rather than the value of the units they sell. 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 research agenda.

An interesting question that follows from the firm-size interpretation is then what are the determinants of firm-size? Controlling for productivity also controls for firm-size determinants such as managerial ability so variation in firm-size might be limited in our model. There is a large and historical literature on the determinants of firm-size. Kumar *et al.* (1999) survey some of the theories and test them using data. We offer a possible explanation below, but we focus on whether there is heterogeneity in firm-size and its consequences.

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 idea that input choices can impact firm's production beyond the conditional mean has important consequences 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. These models are commonly applied to studying 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.

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.

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.

3.1 Identification

3.1.1 Production Function in Levinsohn and Petrin (2003)

We follow LP in the usual set of assumptions on timing of input choices and scalar unobservability.

Assumption 3.1

(a) The production function $y_{it} = \beta_k(\eta_{it})k_{it} + \beta_l(\eta_{it})l_{it} + \omega_{it}$ *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 exogenously 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 $\mathbb{E}[\xi_{it}|\mathcal{I}_{it-1}] = 0$ and $P[\xi_{it} \leq G^{-1}(\tau_\xi)|\mathcal{I}_{it-1}] = \tau_\xi$, where $G(\cdot)$ is the CDF of ξ_{it}

(d) Firms accumulate capital according to

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

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

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

Given these Assumption (3.1)(e), 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) and (b) we have the following identification condition for the first stage:

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

Here we make a distinction between τ (rank of ex-post shock) and τ_ξ (rank of innovation shock to productivity) since there are multiple unobservables in our model. The equation in (15) is a semi-parametric partially linear quantile regression model which can be consistently estimated using local polynomials (Lee, 2003), smoothing splines, Koekner *et al.* (1994), or by sieve minimum distance (SMD) (Chen and Pouzo, 2009). Let $\hat{y}_{it} = y_{it} - \hat{\beta}_l(\tau)l_{it}$ denote output net of the estimated labor contribution. In the second stage we have

$$\hat{y}_{it} = \beta_k(\eta_{it})k_{it} + \omega_{it} \tag{17}$$

The conditional quantile of equation (17) can be obtained by replacing ω_{it} by its conditional quantile in an iterated approach:

$$Q_{y_{it}|\mathcal{I}_{it-1}}(\tau, k_{it}, Q_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1})) = \beta_k(\tau)k_{it} + Q_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1}), \tag{18}$$

where we have used the assumption of exogeneity of the Markov process for productivity $Q_{\omega_{it}|\mathcal{I}_{it-1}} = Q_{\omega_{it}|\omega_{it-1}}$. The conditional quantiles of productivity according to the model we specify in Assump-

tion (3.1)(c) is given by

$$Q_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1}) = g(\omega_{it-1}) + G^{-1}(\tau_\xi) \tag{19}$$

In our model, productivity can be written as a function of first stage estimates and the capital elasticity

$$\begin{aligned} \omega_{it} &= \Phi(k_{it}, m_{it}, \eta_{it}) - \beta_k(\eta_{it})k_{it} = E[\omega_{it}|\omega_{it-1}, \eta_{it-1}] + \xi_{it} \\ &= g(\Phi(k_{it-1}, m_{it-1}, \eta_{it-1}) - \beta_k(\eta_{it-1})k_{it-1}) + \xi_{it} \end{aligned} \tag{20}$$

The production shock η_{it} enters the productivity equation in a complicated way, we proceed to integrate equation (20) with respect to the marginal distribution of η_{it}. Using the law of iterated expectations then yields

$$\omega_{it} = \bar{\Phi}(k_{it}, m_{it}) - \bar{\beta}_k k_{it} = g(\bar{\Phi}(k_{it-1}, m_{it-1}) - \bar{\beta}_k k_{it-1}) + \xi_{it}, \tag{21}$$

where $\bar{\Phi}(k_{it}, m_{it}) = \int_0^1 \Phi(k_{it}, m_{it}, \tau)d\tau$ and $\bar{\beta}_k = \int_0^1 \beta_k(\tau)d\tau$. These objects are identified by the conditional mean counterparts of our random coefficient model

$$\mathbb{E}[y_{it}|\mathcal{I}_{it}] = \mathbb{E}[\beta_l(\eta_{it})l_{it} + \Phi(k_{it}, m_{it}, \eta_{it})|\mathcal{I}_{it}] = \bar{\beta}_l l_{it} + \bar{\Phi}(k_{it}, m_{it}) \tag{22}$$

and in the second stage

$$\begin{aligned} \mathbb{E}[\hat{y}_{it}|\mathcal{I}_{it-1}] &= \mathbb{E}[\beta_k(\eta_{it})k_{it} + g(\Phi(k_{it-1}, m_{it-1}, \eta_{it-1}) - \beta_k(\eta_{it-1})k_{it-1}) + \xi_{it}] \\ &= \bar{\beta}_k k_{it} + g(\bar{\Phi}(k_{it-1}, m_{it-1}) - \bar{\beta}_k k_{it-1}) \end{aligned} \tag{23}$$

which identifies $\bar{\beta}_k$ and $g(\cdot)$. The conditional mean conditions in this model are used to identify $\bar{\Phi}$ and $\bar{\beta}_k$ similar to the LP identification strategy. To identify $G^{-1}(\tau_\xi)$ we use

$$P(\xi_{it} \leq G^{-1}(\tau_\xi)|\mathcal{I}_{it-1}) = P(\xi_{it} \leq G^{-1}(\tau_\xi)) = \tau_\xi \tag{24}$$

from Assumption (3.1)(c). Therefore the conditional quantile $Q_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1}))$ is identified from (22), (23) and (24). Let $\hat{Q}_{\omega_{it}|\omega_{it-1}}$ denote a consistent estimator of this function. Plugging this into to the second stage conditional quantile in our model yields

$$Q_{y_{it}|\mathcal{I}_{it-1}}(\tau, k_{it}, Q_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1})) = \beta_k(\tau)k_{it} + \hat{Q}_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1}),$$

Let $\tilde{y}_{it} = \hat{y}_{it} - \hat{Q}_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1})$. Using the conditional moment representation of the above equation we have

$$\mathbb{E}[\Psi_\tau(\tilde{y}_{it} - \beta_k(\tau)k_{it})|\mathcal{I}_{it-1}] = 0 \tag{25}$$

where $\Psi_\tau(u) = \tau - \mathbb{1}\{u < 0\}$. The unconditional moments are then

$$\mathbb{E}[Z_{it}\Psi_\tau(\tilde{y}_{it} - \beta_k(\tau)k_{it})] = 0 \tag{26}$$

where Z_{it} is a subset of instrumental variables from $\mathcal{I}_{it-1}$. When $Z_{it} = k_{it}$, equation (26) is the first-order condition of the quantile regression estimator of $\beta_k(\tau)$. Over-identification conditions can be pursued, which require additional assumptions from the IVQR literature. For example, Chernozhukov and Hansen (2005) give strong conditions for global identification of their model using the global version of a full rank assumption. We rely on a much weaker full rank condition that is sufficient for local identification of $\beta_k(\tau)$ provided by de Castro *et al.* (2018).

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:

$$y_{it} = \beta_k(\eta_{it})k_{it} + \beta_l(\eta_{it})l_{it} + m_t^{-1}(k_{it}, l_{it}, m_{it}) = \Phi(k_{it}, l_{it}, m_{it}, \eta_{it}), \tag{27}$$

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)

$$P\big(y_{it} \leq \Phi(k_{it}, l_{it}, m_{it}; \tau)\big|\mathcal{I}_{it}\big) = \tau \tag{28}$$

so that 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). However, since we use a conditional mean version of Φ to estimate productivity it would suggest that estimation of $\Phi(\cdot, \tau)$ is not actually needed, instead we would use

$$\mathbb{E}[y_{it}|\mathcal{I}_{it}] = \mathbb{E}[\Phi(k_{it}, l_{it}, m_{it}, \eta_{it})|\mathcal{I}_{it}] = \bar{\Phi}(k_{it}, l_{it}, m_{it}) \tag{29}$$

and in the second stage

$$\begin{aligned}\mathbb{E}[y_{it}|\mathcal{I}_{it-1}] &= \mathbb{E}[\beta_k(\eta_{it})k_{it} + \beta_l(\eta_{it})l_{it} + g(\Phi(k_{it-1}, l_{it-1}, m_{it-1}, \eta_{it-1}) - \beta_k(\eta_{it-1})k_{it-1} - \beta_l(\eta_{it-1})l_{it-1}) + \xi_{it}] \\ &= \bar{\beta}_k k_{it} + \bar{\beta}_l l_{it} + g(\bar{\Phi}(k_{it-1}, l_{it-1}, m_{it-1}) - \bar{\beta}_k k_{it-1} - \bar{\beta}_l l_{it-1}),\end{aligned} \tag{30}$$

which identifies $\bar{\beta}_l, \bar{\beta}_k$ and $g(\cdot)$. The same identification strategy for $G^{-1}(\tau_\xi)$ is given by equation (24). The second stage conditional quantile is then

$$Q_{y_{it}|\mathcal{I}_{it-1}}(\tau, k_{it}, l_{it}, Q_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1})) = \beta_k(\tau)k_{it} + \beta_l(\tau)l_{it} + \hat{Q}_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1}),$$

Let $\hat{y}_{it} = y_{it} - \hat{Q}_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1})$. Using the conditional moment representation of the above equation we have

$$\mathbb{E}[\Psi_\tau(\hat{y}_{it} - \beta_k(\tau)k_{it} - \beta_l(\tau)l_{it})|\mathcal{I}_{it-1}] = 0 \tag{31}$$

The unconditional moments are then

$$\mathbb{E}[Z_{it}\Psi_\tau(\hat{y}_{it} - \beta_k(\tau)k_{it} - \beta_l(\tau)l_{it})] = 0 \tag{32}$$

3.2 Estimation

We abbreviate our estimator as QLP since it is the quantile version of the LP method. We use a similar meaning for the abbreviation QACF. We discuss how to estimate the QLP production function in two stages. We omit the estimation steps for the QACF production function since they are similar to the QLP strategy minus the first step estimator.

First Stage

Recall, in the first stage we have the identification condition

$$P\big(y_{it} \leq \beta_l(\tau)l_{it} + \Phi(k_{it}, m_{it}; \tau)\big|\mathcal{I}_{it}\big) = \tau \tag{33}$$

which yields

$$\mathbb{E}\big[\mathbb{1}\{y_{it} \leq \beta_l(\tau)l_{it} + \Phi(k_{it}, m_{it}; \tau)\} - \tau\big|\mathcal{I}_{it}\big] = 0 \tag{34}$$

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 review 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

$$\mathbb{E}[\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi(x_{it}))|\mathcal{I}_{it}] = 0 \tag{35}$$

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

$$(\hat{\beta}_l, \hat{\Phi}) = \underset{\alpha \in (\Theta \times \mathcal{H}_{k(n)})}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\mathbb{E}}[\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi)|\mathcal{I}_{it}]' \hat{\Sigma}_1^{-1} \hat{\mathbb{E}}[\Lambda_1(y_{it}, l_{it}; \beta_l, \Phi)|\mathcal{I}_{it}] \tag{36}$$

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 empirical application we use a 3rd order polynomial with interactions and estimate the first stage parameters using simple weighted linear quantile regression.

Second Stage

Once estimates of $\beta_l(\tau)$ are obtained, construct the new response variable as output net of the labor contribution $\hat{y}_{it} = y_{it} - \hat{\beta}_l(\tau) l_{it}$. Estimates of the conditional mean counterparts of $\Phi(k_{it}, l_{it}, \eta_{it})$ and $\beta_k(\eta_{it})$, $\bar{\Phi}(k_{it}, l_{it})$, $\bar{\beta}_k$ and $g(\cdot)$ can be obtained using the LP estimator described in Section 2. Then estimates of $G^{-1}(\tau_\xi)$ can be obtained from the sample τ_ξ-th quantiles of $\hat{\xi}_{it} = \omega_{it} - \hat{g}(\omega_{it-1})$. In practice, the choice of τ_ξ does not matter for estimation of $\beta_k(\tau)$ so we choose $\tau_\xi = 0.5$ for every τ. Then the estimates for the conditional quantiles of productivity are given by $\hat{Q}_{\omega_{it}|\omega_{it-1}}(\tau_\xi, \omega_{it-1}) = \hat{g}(\hat{\bar{\Phi}}(k_{it-1}, m_{it-1}) - \hat{\bar{\beta}}_k k_{it-1}) + \hat{G}^{-1}(\tau_\xi)$ which is subtracted from $\hat{y}_{it}$ to obtain $\tilde{y}_{it}$.

Estimation of $\beta_k(\tau)$ using the sample moments corresponding to equation (26) is infeasible due to the indicator function inside Ψ_τ. Therefore we smooth this function using the method of Kaplan and Sun (2016) and consider over-identification conditions using a GMM criterion function from de Castro *et al.* (2018). The smoothed sample moments are then given by

$$\hat{M}(\beta_k(\tau)) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Z_{it} \left[\tilde{\Psi}_\tau \left(\frac{\tilde{y}_{it} - \beta_k(\tau) k_{it}}{h_n} \right) \right] \tag{37}$$

where h_n is a bandwidth sequence and $\tilde{\Psi}_\tau(u) = \tau - \tilde{\mathbb{1}}\{u < 0\}$ is a smoothed version of $\Psi_\tau(u)$. A version of the smoothed indicator function has been used by Horowitz (1998), Whang (2006), and Kaplan and Sun (2016) which is given by

$$\tilde{\mathbb{1}}(u) = \mathbb{1}\{-1 \leq u \leq 1\}[0.5 + \frac{105}{64}(u - \frac{5}{3}u^3 + \frac{7}{5}u^5 - \frac{3}{7}u^7)] + \mathbb{1}\{u > 1\}$$

Then $\beta_k(\tau)$ can be found by minimizing the GMM criterion function with smoothed sample mo-

ments

$$\hat{\beta}_k(\tau) = \text{argmin}\, \hat{M}(\beta_k(\tau))' \hat{W} \hat{M}(\beta_k(\tau))$$

where $\hat{W}$ is the weighting matrix. The choice of an optimal weighting matrix is discussed by Kaplan and Sun (2016) for the linear IVQR model with iid data and by de Castro *et al.* (2018) for the nonlinear IVQR model with iid and weakly dependent data. In practice, we use the weighting matrix from Kaplan and Sun (2016) which is an estimate of $[\tau(1-\tau)\mathbb{E}[Z_{it}Z_{it}']]^{-1}$. However, an optimal weighting matrix in this step should reflect the noise in the estimates of the previous step. Asymptotic standard errors could be obtained using Ackerberg *et al.* (2014), however we do not undertake this task here. Instead we use a nonparametric bootstrap with re-centering of the sample moments in the second stage with over-identification to estimate the standard errors.

4 Monte Carlo Experiments

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

$$y_{it} = \beta_0 + \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + (\gamma_0 + \gamma_k k_{it} + \gamma_l l_{it})\eta_{it} \tag{38}$$

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 $\eta_{it} \sim Laplace(0, 0.1)$.

Following the data generating process in Ackerberg *et al.* (2015), 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 the 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 $[\tau(1-\tau)\mathbb{E}[Z_{it}Z_{it}']]^{-1}$ and we set $h = 0.1$ for simplicity.

Table 1 provides estimates of the bias and MSE for DGP 1 and DGP 2. They are both very small and decrease as the estimates approach $\tau = 0.5$. We plot the MSE of our estimator as well as the MSE of the LP estimator compared to the true values of $\beta_k(\tau)$ and $\beta_l(\tau)$ in Figure 1. 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. It is clear from this plot that our estimator does relatively well at capturing the heterogeneity in our model. Lastly we test whether our estimates control for unobserved productivity by comparing our estimates to standard

quantile regression estimates without controlling for productivity. Figure 2 plots estimates of these difference along with their 95% confidence intervals for each τ. The plots show significant differences in these estimates with capital being underestimated and labor being overestimated in the presence of the simultaneity bias.

Table 1: Bias and MSE

		Capital		Labor	
DGP	τ	Bias	MSE	Bias	MSE
1	0.10	-0.0017	0.0013	0.0059	0.0009
	0.25	0.0008	0.0012	0.0015	0.0000
	0.50	0.0020	0.0012	0.0000	0.0000
	0.75	0.0032	0.0012	-0.0015	0.0000
	0.90	0.0057	0.0013	-0.0049	0.0003
2	0.10	-0.0057	0.0022	0.0056	0.0001
	0.25	-0.0025	0.0023	0.0026	0.0001
	0.50	-0.0010	0.0021	0.0000	0.0000
	0.75	0.0015	0.0021	-0.0026	0.0000
	0.90	0.0037	0.0021	-0.0056	0.0001

Figure 1: Simulated precision of QLP estimators of $\beta_k(\tau)$ and $\beta_l(\tau)$s. Dotted line is LP estimator.

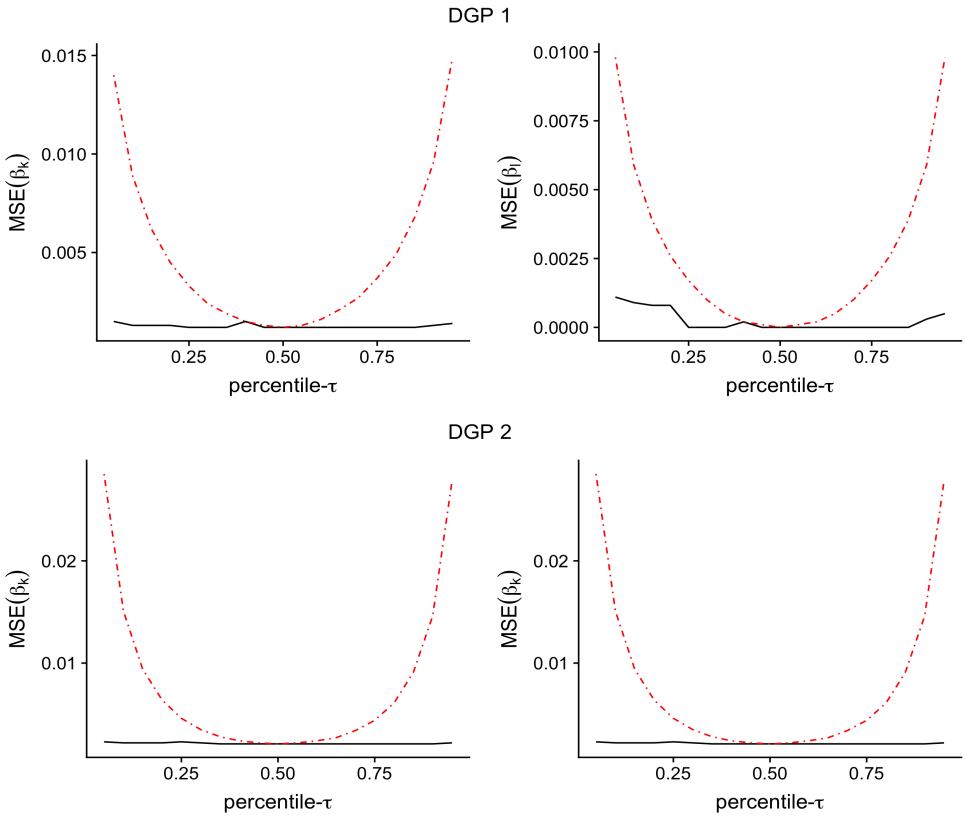
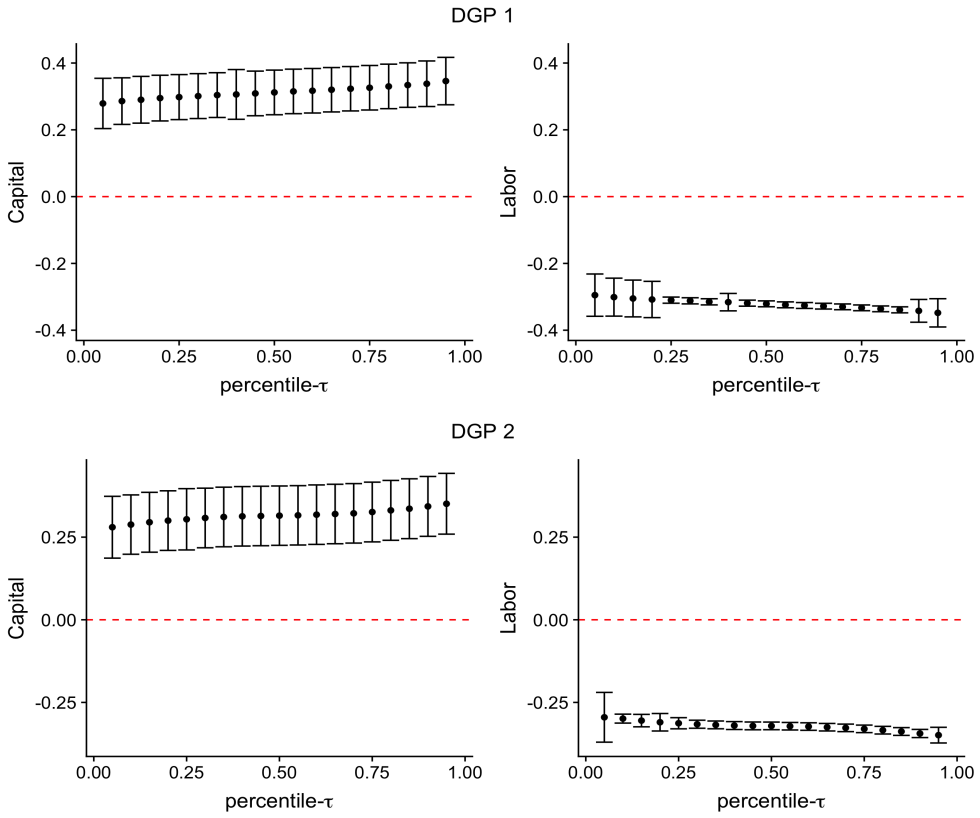


Figure 2: Difference between QLP estimators and QR estimates of $\beta_k(\tau)$ and $\beta_l(\tau)$s.



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 [6]. For each country we examine estimates across different manufacturing industries as well as how these estimates have changed over time. 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 and Φ with a 3rd degree polynomial with interactions in capital and materials. To estimate capital, we use the smoothed GMM criterion function mentioned earlier with instruments $k_{it}, k_{it-1}, l_{it-1}$ and m_{it-1}. We use the plug-in method of Kaplan and Sun (2016) for bandwidth selection. We initialize the starting values for the non-linear search using quantile regression estimates of $\beta_k(\tau)$ from the exactly identified model. 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 re-center sample moments in each iteration.

[6] 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 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 2: 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

Estimates of the capital elasticity are slightly increasing in the firm-size distribution in every industry except for NAICS 31 where estimates drop dramatically after $\tau = 0.1$ then remain flat. The estimates for labor elasticity corresponding to NAICS 31 and 32 are decreasing. For NAICS 33 and the entire sample, the labor elasticity is an inverse U-shape; it increases quickly for low τ then flattens after $\tau = 0.5$. In NAICS 33 and the entire sample only the labor estimates are significantly different from the LP estimate after $\tau = 0.25$. In each industry we compare the dif-

ference 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 capital elasticity in NAICS 31 for very small firms and medium-sized/large firms in NAICS 32. This may suggest that in these smaller industries that capital usage for these types of firms responds less to productivity shocks and there is not much difference between capital elasticity and the size of the firm. For NAICS 33 and the entire sample, capital usage responds to productivity for all firm sizes, but after controlling for productivity, there is not much variation in capital elasticity between firms.

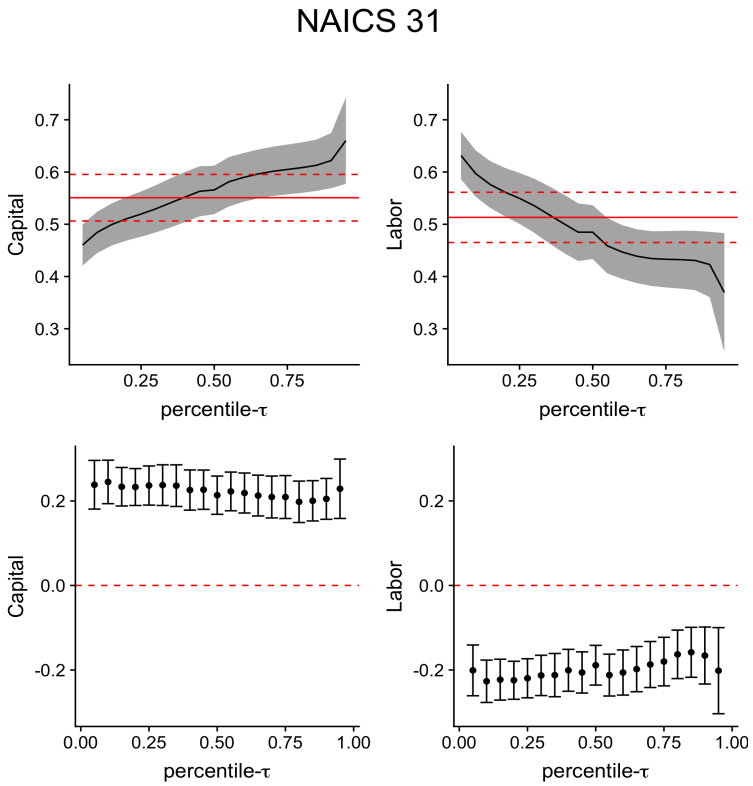


Figure 3: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

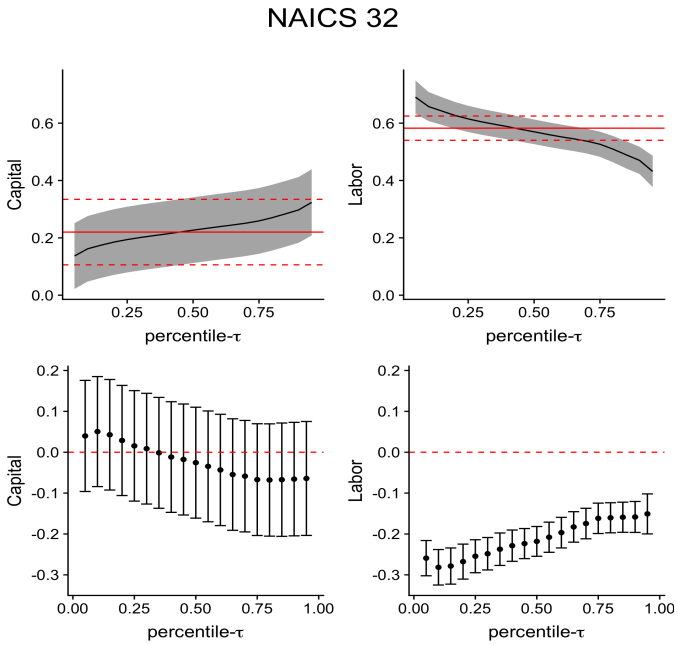


Figure 4: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

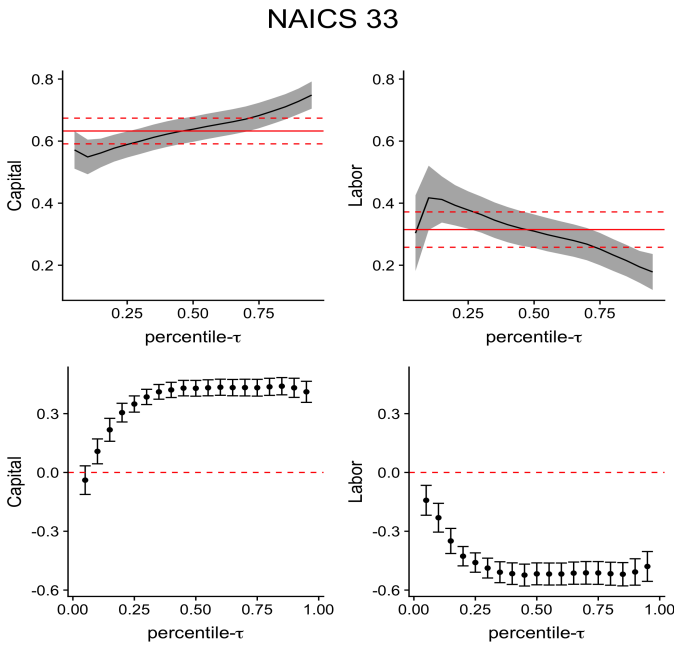


Figure 5: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

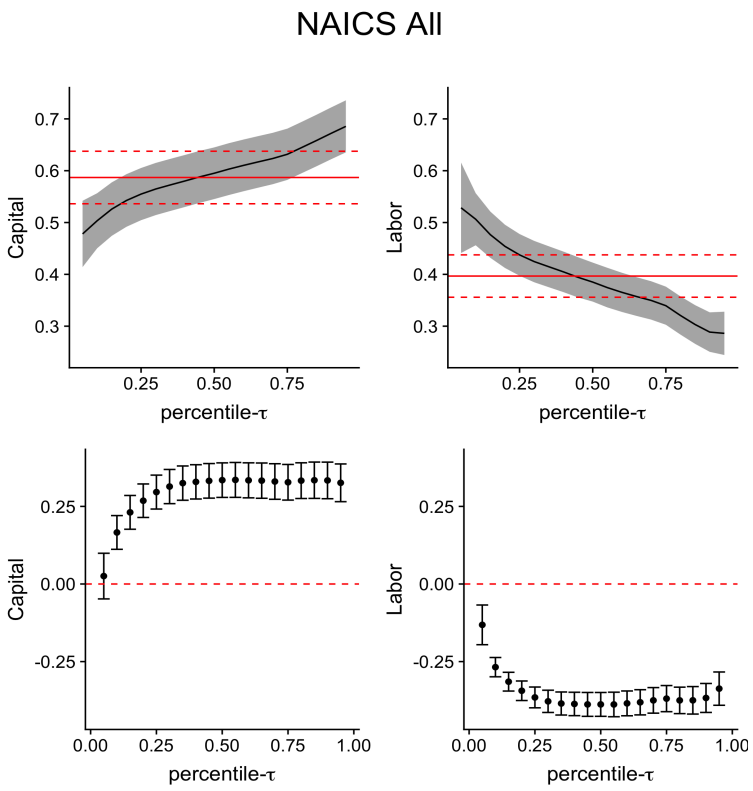


Figure 6: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

We use the estimates from the output elasticities to construct measures of returns to scale and capital intensity in Table 3. The results for returns to scale are puzzling as they are all significantly different from constant returns to scale. We can see in NAICS 32 that returns to scale are generally decreasing in firm-size whereas this relationship is increasing in NAICS 33 and the entire sample. 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. Aside from NAICS 31, each industry has increasing capital intensity. This result is consistent with previous findings such as Holmes and Mitchell (2008), Kumar *et al.* (1999) and Dermirer (2020).

We also use our quantile production function estimates to construct measures of firm level

Table 3: Coefficient Estimates and Standard Errors for US Manufacturing Firms

		Capital		Labor		Returns to Scale		Capital Intensity	
NAICS	τ	Coef.	s.e	Coef.	s.e	Coef.	s.e	Coef.	s.e
31	0.10	0.485	0.0240	0.597	0.0270	1.081	0.0224	0.812	0.0654
	0.25	0.520	0.0263	0.550	0.0296	1.069	0.0219	0.945	0.0873
	0.50	0.566	0.0282	0.485	0.0313	1.051	0.0211	1.167	0.1196
	0.90	0.622	0.0322	0.423	0.0380	1.045	0.0217	1.471	0.2074
32	0.10	0.162	0.0694	0.658	0.0308	0.820	0.0747	0.246	0.1055
	0.25	0.194	0.0697	0.615	0.0278	0.809	0.0740	0.315	0.1135
	0.50	0.227	0.0697	0.570	0.0261	0.797	0.0739	0.398	0.1228
	0.90	0.297	0.0697	0.470	0.0289	0.768	0.0743	0.632	0.1547
33	0.10	0.549	0.0339	0.417	0.0629	0.966	0.0344	1.316	0.2873
	0.25	0.589	0.0255	0.378	0.0366	0.967	0.0195	1.559	0.2051
	0.50	0.638	0.0247	0.310	0.0328	0.948	0.0170	2.061	0.2832
	0.90	0.728	0.0250	0.195	0.0312	0.923	0.0158	3.739	0.7235
All	0.10	0.503	0.0323	0.506	0.0305	1.010	0.0262	0.994	0.1071
	0.25	0.555	0.0306	0.438	0.0245	0.992	0.0252	1.268	0.1207
	0.50	0.595	0.0303	0.385	0.0229	0.980	0.0250	1.545	0.1467
	0.90	0.672	0.0304	0.289	0.0232	0.961	0.0253	2.329	0.2567

productivity which we define as

$$\hat{w}_{it,\tau} = \exp(y_{it} - \hat{\beta}_k(\tau)k_{it} - \hat{\beta}_l(\tau)l_{it}) \tag{39}$$

We use these measures to compare productivity growth over time to LP estimates 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 $\tau = 0.25$. This suggests that there is significant heterogeneity in productivity across the conditional firm-size distribution that conditional mean estimates of productivity such as LP may not capture.

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 8 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. In the beginning of the sample

period, large firms have larger estimates of capital elasticity than smaller firms which are similar in magnitude. These estimates are roughly the same until 1990 when the estimates increase sharply for very small firms but then falls at the year 2000.

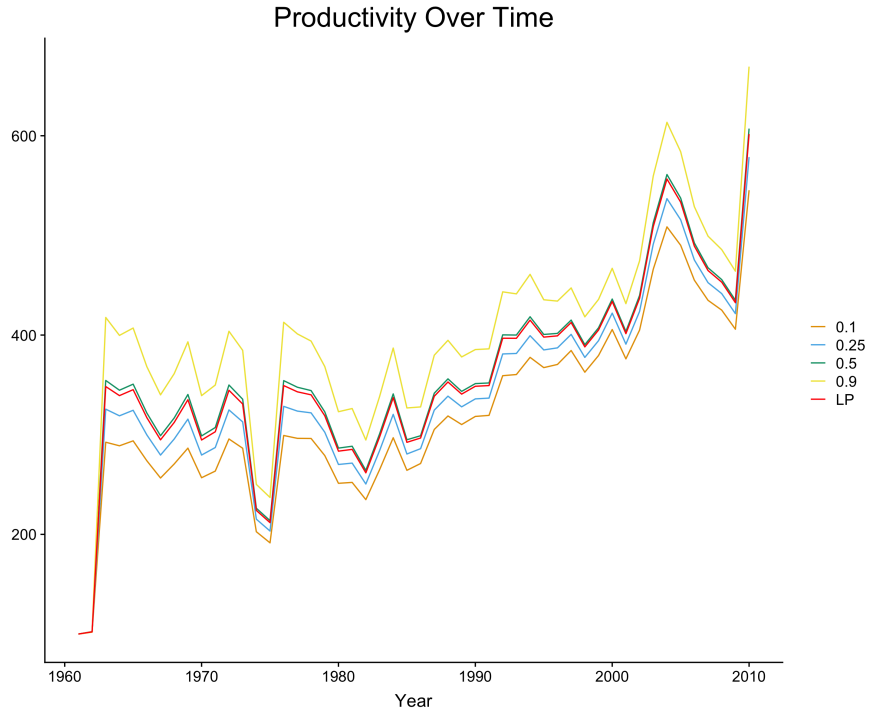


Figure 7: Estimated average productivity over time for the US. Base productivity in 1961 is set to 100.

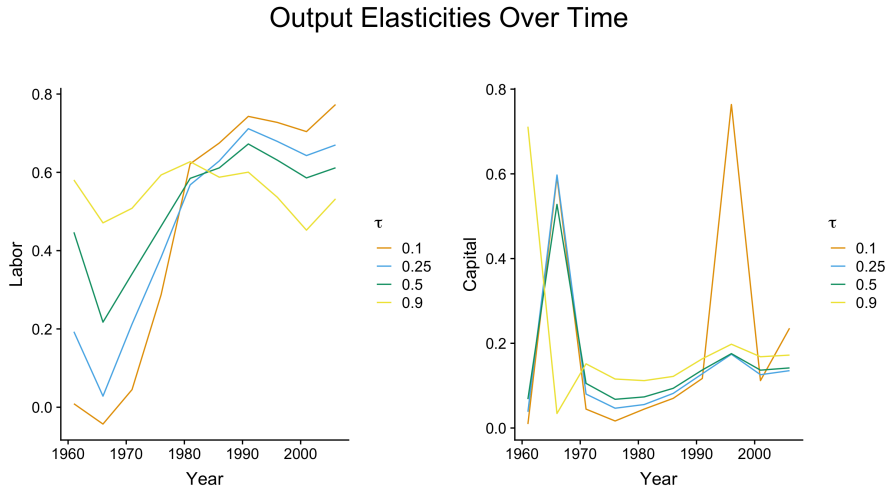


Figure 8: Estimated values of production function coefficients over time estimated at 5 year intervals

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 4.

Figures 9, 10, and 11 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 it suggests that much of the heterogeneity in the labor estimates come from productivity rather than firm-size. The estimates for capital elasticity are increasing in each industry. In the combined sample, these estimates increase then fall sharply at $\tau = 0.75$ then increases sharply for much larger firms. All of the capital estimates are significantly different from the LP estimates. From Table 5, our estimates of returns to scale are more reasonable compared to the US estimates aside from small firms in ISIC 311 who experience decreasing returns to scale. In this industry, returns to scale increase with firm-size.

Table 4: 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

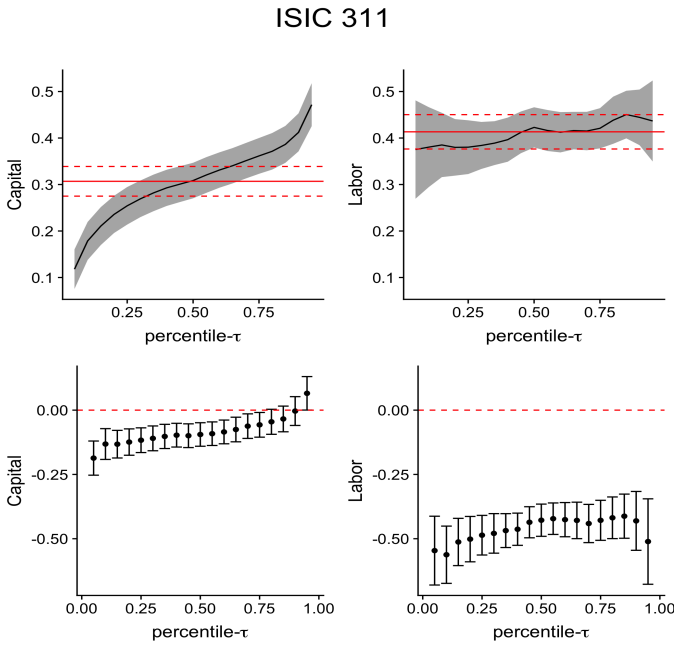


Figure 9: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

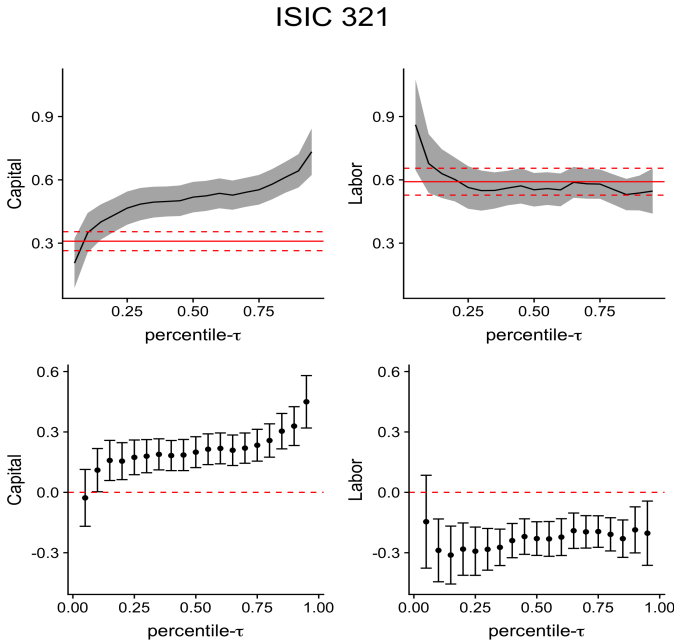


Figure 10: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

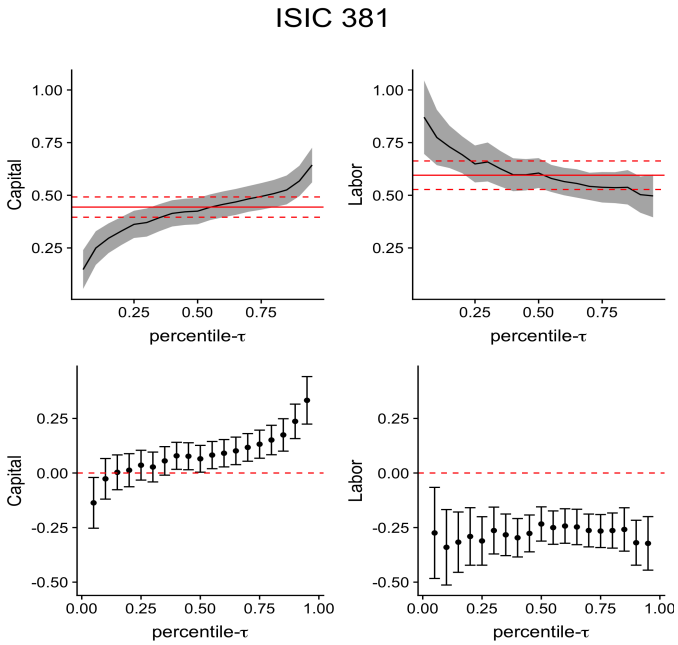


Figure 11: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

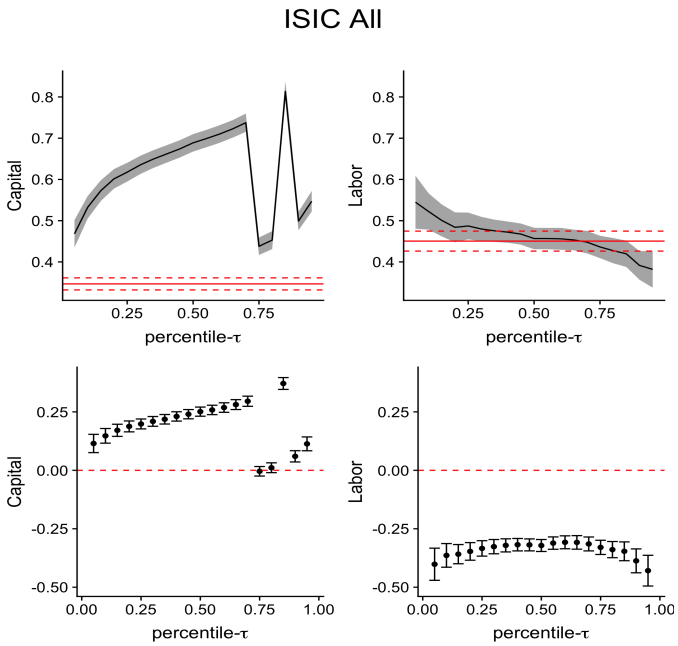


Figure 12: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

Table 5: Coefficient Estimates and Standard Errors for Chilean Manufacturing Firms

		Capital		Labor		Returns to Scale		Capital Intensity	
ISIC	τ	Coef.	s.e	Coef.	s.e	Coef.	s.e	Coef.	s.e
311	0.10	0.179	0.0248	0.380	0.0523	0.559	0.0486	0.470	0.1071
	0.25	0.254	0.0246	0.380	0.0351	0.635	0.0357	0.668	0.0929
	0.50	0.309	0.0232	0.423	0.0263	0.731	0.0310	0.730	0.0807
	0.90	0.412	0.0248	0.444	0.0364	0.857	0.0354	0.927	0.1114
381	0.10	0.250	0.0485	0.774	0.0800	1.024	0.0570	0.322	0.1019
	0.25	0.362	0.0393	0.649	0.0533	1.011	0.0460	0.558	0.1029
	0.50	0.425	0.0379	0.606	0.0431	1.030	0.0405	0.701	0.1175
	0.90	0.568	0.0439	0.503	0.0523	1.071	0.0450	1.129	0.2155
321	0.10	0.349	0.0571	0.677	0.0847	1.026	0.0559	0.516	0.1375
	0.25	0.466	0.0478	0.564	0.0614	1.030	0.0455	0.827	0.1424
	0.50	0.518	0.0440	0.553	0.0484	1.071	0.0408	0.937	0.1403
	0.90	0.643	0.0481	0.538	0.0502	1.180	0.0440	1.195	0.1932
All	0.10	0.532	0.0166	0.522	0.0267	1.055	0.0171	1.020	0.0799
	0.25	0.618	0.0139	0.487	0.0197	1.105	0.0143	1.268	0.0725
	0.50	0.689	0.0128	0.457	0.0158	1.145	0.0127	1.508	0.0720
	0.90	0.499	0.0136	0.391	0.0211	0.891	0.0183	1.275	0.0935

Figure 13 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 then increases for the rest of the sample period. The LP estimates show higher productivity than the larger firms at $\tau = 0.9$. Figure 14 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 estimates for capital elasticities follow similar trends however there is no noticeable ranking of the magnitude of the elasticities across firm-size.

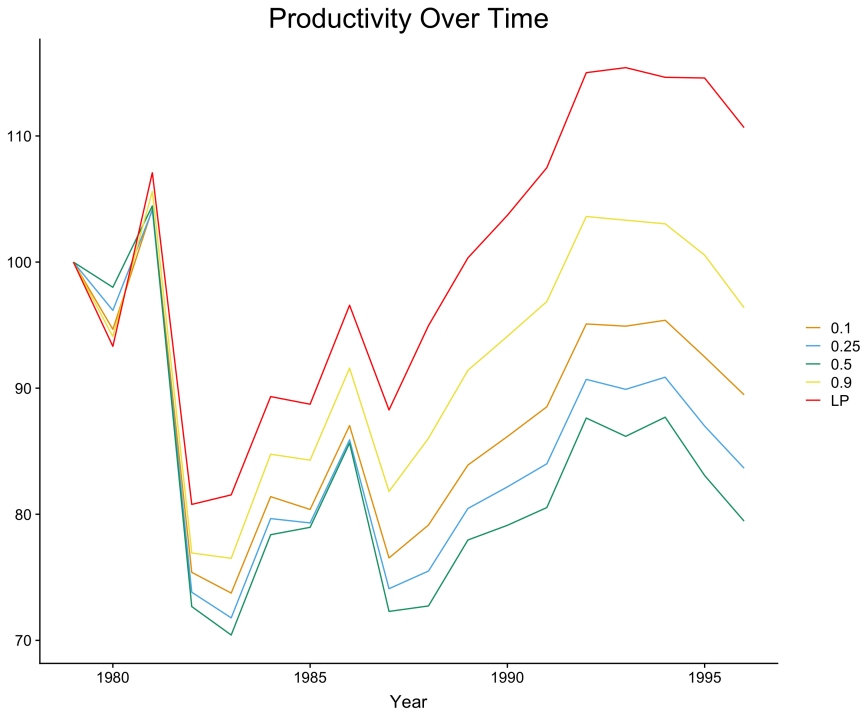


Figure 13: Estimated average productivity over time for Chile. Base productivity in 1979 is set to 100.

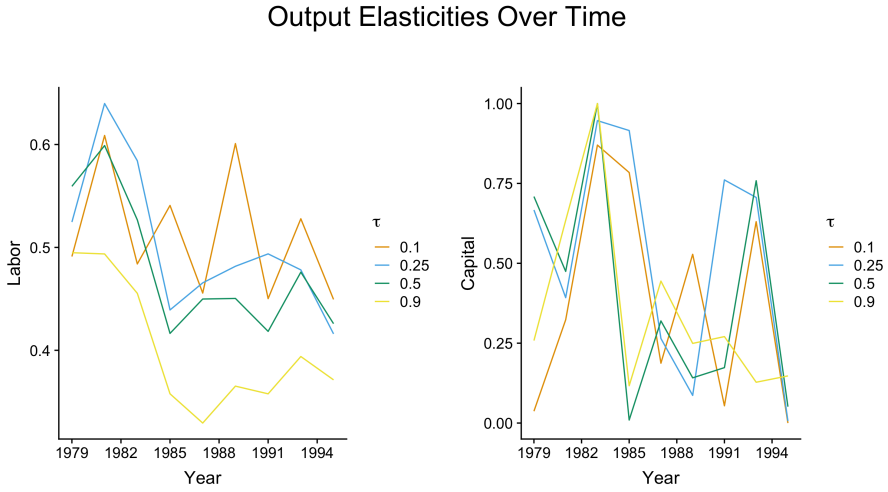


Figure 14: Estimated values of production function coefficients over time estimated at 2 year intervals

5.3 Colombian Manufacturing

This data comes from the Colombian manufacturing census conducted by the Departamento Administrativo Nacional de Estadistica. The sample is collected between 1977 and 1991. 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 6.

Table 6: 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 15, 16, and 17 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. In ISIC 322, both estimates show significant differences from the LP estimate, however the capital elasticity appears to be unreasonably high. In ISIC 381 there are only differences in the labor estimates. With the combined sample both QLP estimates of capital and labor are significantly different from LP and QR estimates.

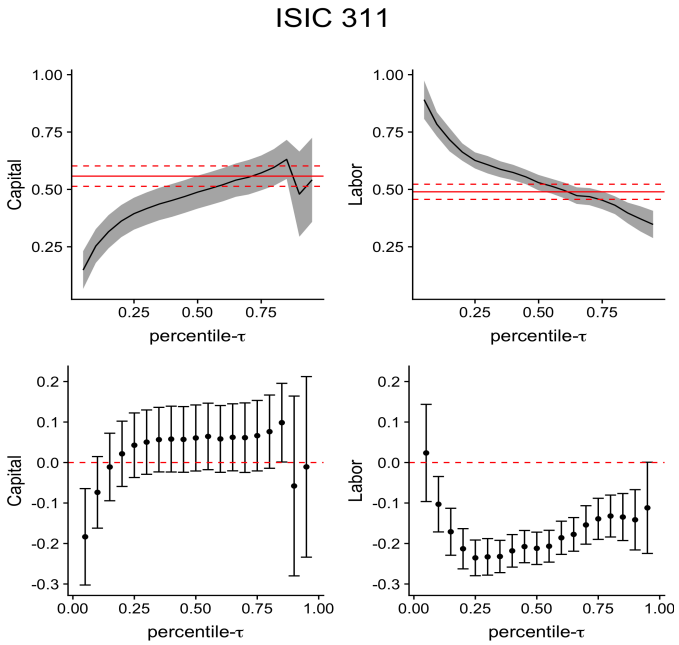


Figure 15: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

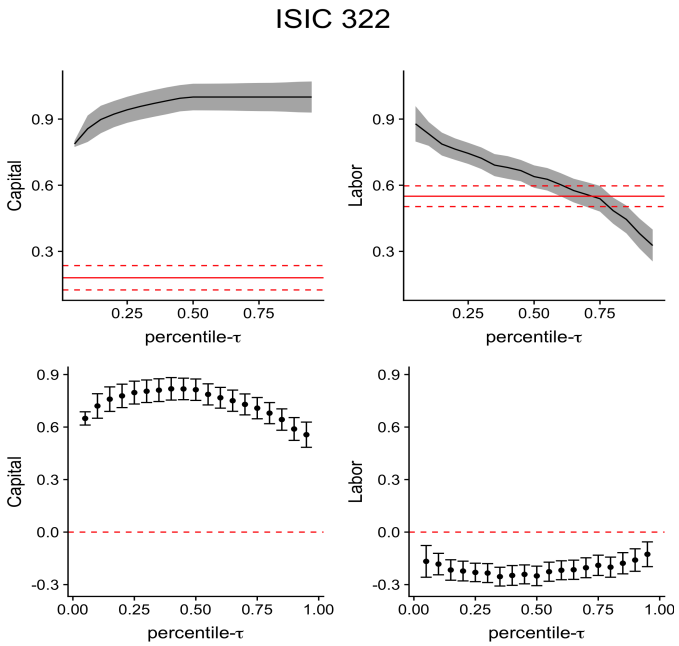


Figure 16: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

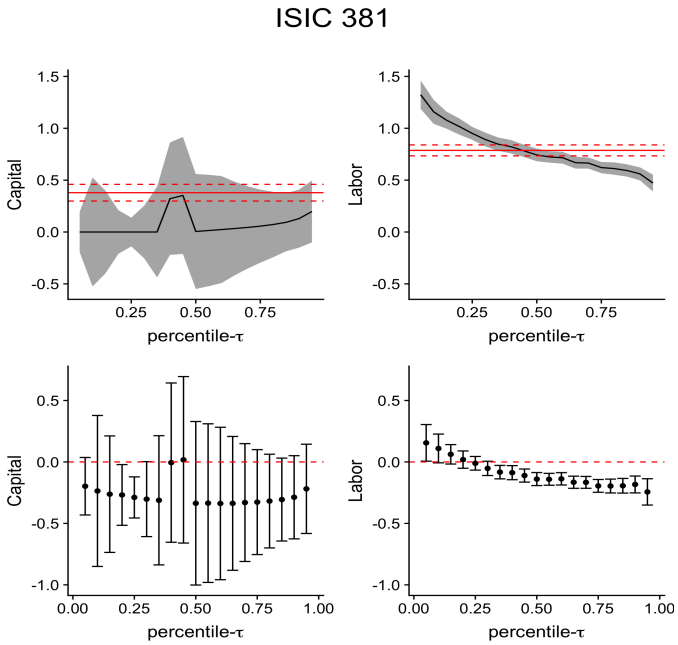


Figure 17: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

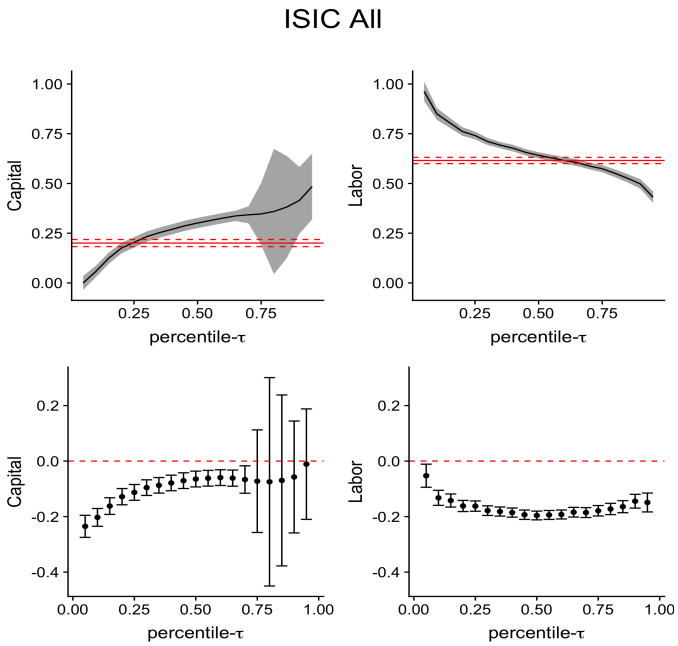


Figure 18: Top row: Estimated values of production function coefficients and their point-wise 90% confidence interval. Bottom row: Difference between QLP and quantile regression estimates and their 95% confidence intervals.

Using these estimates we construct measures of returns to scale and capital intensity for each industry in Table 7. Most firms experience constant returns to scale or significant increasing returns in ISIC 322. We observe that returns to scale are decreasing in firm-size in each industry, but not the combined sample. Capital intensity increases in firm-size for each industry and the combined sample.

Table 7: Coefficient Estimates and Standard Errors for Colombian Manufacturing Firms

		Capital		Labor		Returns to Scale		Capital Intensity	
ISIC	τ	Coef.	s.e	Coef.	s.e	Coef.	s.e	Coef.	s.e
311	0.10	0.254	0.0448	0.784	0.0318	1.038	0.0418	0.323	0.0662
	0.25	0.394	0.0423	0.626	0.0214	1.021	0.0416	0.630	0.0773
	0.50	0.488	0.0437	0.529	0.0205	1.018	0.0431	0.922	0.1026
	0.90	0.479	0.1128	0.372	0.0333	0.851	0.1159	1.290	0.3688
322	0.10	0.856	0.0365	0.833	0.0329	1.689	0.0375	1.028	0.0455
	0.25	0.942	0.0359	0.745	0.0292	1.687	0.0338	1.265	0.0534
	0.50	1.000	0.0367	0.639	0.0309	1.639	0.0335	1.564	0.0728
	0.90	1.000	0.0421	0.381	0.0413	1.381	0.0372	2.625	0.2150
381	0.10	0.000	0.3189	1.159	0.0703	1.159	0.2928	0.000	0.2930
	0.25	0.000	0.0843	0.951	0.0379	0.951	0.0911	0.000	0.1019
	0.50	0.005	0.3367	0.743	0.0369	0.748	0.3414	0.007	0.4534
	0.90	0.129	0.1702	0.560	0.0383	0.689	0.1753	0.231	0.3457
All	0.10	0.058	0.0176	0.850	0.0179	0.907	0.0167	0.068	0.0240
	0.25	0.204	0.0160	0.741	0.0118	0.945	0.0153	0.275	0.0266
	0.50	0.301	0.0156	0.641	0.0095	0.943	0.0153	0.470	0.0307
	0.90	0.415	0.1022	0.498	0.0145	0.913	0.1029	0.834	0.2131

Figure 19 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 smaller than productivity of large firms at $\tau = 0.9$. Figure 20 shows the time trends in output elasticities. The estimates 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 move together aside from the very large firms. These estimates have an inverse-U shaped relationship over time.

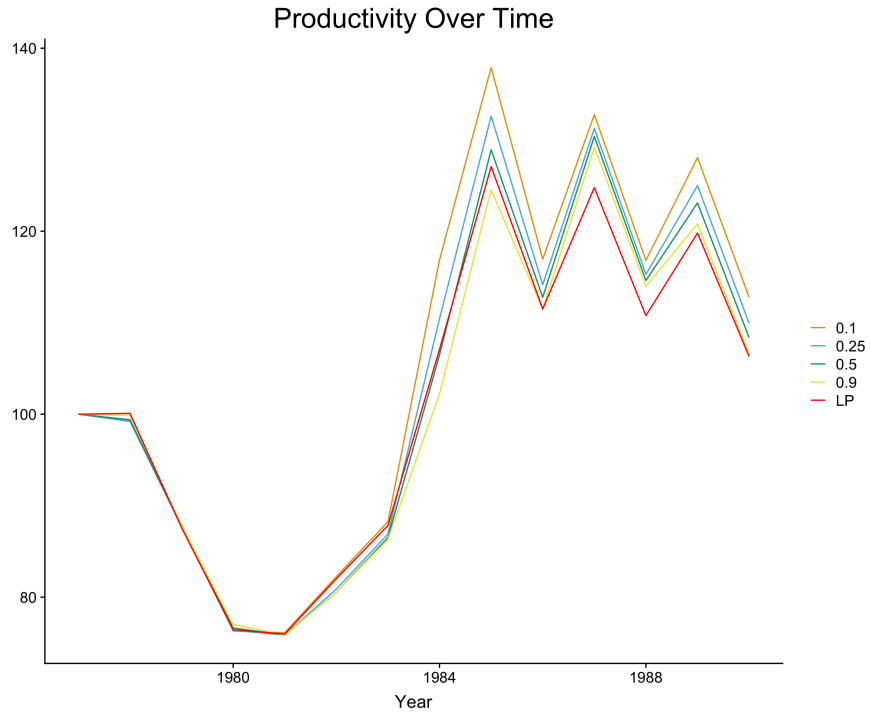


Figure 19: Estimated average productivity over time for Colombia. Base productivity in 1978 is set to 100.

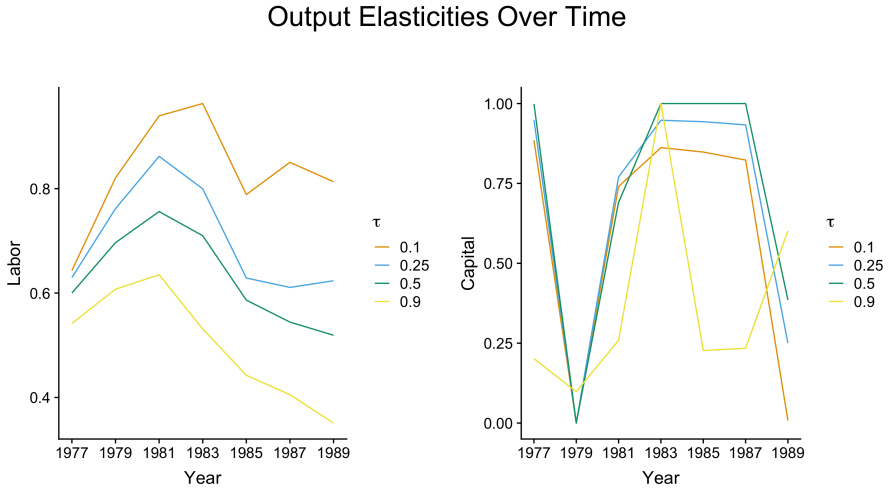


Figure 20: Estimated values of production function coefficients over time estimated at 2 year intervals

6 Conclusions

We propose a method that extends the intermediate input proxy variable approach to estimating the conditional quantiles of firm-size. 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 and second 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 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. Allowing richer distributional effects of productivity would be an interesting extension. This approach also restricts us from examining non-Hicks neutral productivity shocks such as 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., Hahn, J. 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. (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).

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.

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.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, **71** (5), 1405–1441.

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.

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.

Horowitz, J. L. (1998). Bootstrap methods for median regression models. *Econometrica*, **66** (6), 1327.

Imbens and Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, **77** (5), 1481–1512.

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).

— (2007). Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics*, **141** (2), 1131–1158.

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.

Ma, L. and Koenker, R. (2006). Quantile regression methods for recursive structural equation models. *Journal of Econometrics*, **134** (2), 471–506.

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.

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

Whang, Y.-J. (2006). Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory*, **22** (02).