Identifying Market Power in Production Data

Zach Flynn, Amit Gandhi, and James Traina*

Afiniti, University of Pennsylvania, and University of Chicago

February 27, 2019

PRELIMINARY AND INCOMPLETE

Abstract

Production-based estimates of markups require output elasticities for a flexible input, but these elasticities are not identified under the standard assumptions of proxy variable estimators. We show markups are identified given an additional economic restriction: constant returns to scale technology. We present Monte Carlo evidence that ignoring the identification problem, as in prior literature, introduces significant bias in estimated markups. Comparing estimators on US public firm data, we find that our approach is also more robust to testable forms of misspecification. Emerging macroeconomic models imply output and labor share wedges using our markup estimates of 1 and 11 percent, half the size of non-identified estimators.

Emerging macroeconomic research argues that a rise in market power explains a number of important secular trends: the slowdown in output growth, the fall in the labor share and business dynamism, the rise in industrial concentration and corporate profits, among others.[1] Many of these arguments rely on econometric methods that estimate markups using firm optimization conditions. Cost minimization with respect

*We thank Enghin Atalay, John Grigsby, Ken Hendricks, Ivan Kirov, and Alan Sorensen for helpful discussions. We also thank seminar participants at Chicago Booth's Stigler Center Workshop and the ASSA 2019 Meetings for their comments and feedback. Corresponding author address: traina@uchicago.edu.

[1] See Syverson (2019) for a review.

to a flexible[2] input implies:

$$\frac{P}{\text{MC}} = \frac{\text{Revenue}}{\text{Expenditure on Flexible Input}} \times \text{Output Elasticity of Flexible Input} \qquad (1)$$

Hence, if we can estimate the output elasticity of a flexible input from production data, then we can solve for the markup of a cost-minimizing firm. Using the first order conditions of the cost minimization problem to recover markups was first popularized by Hall (1988), and combined with modern proxy variable production estimators[3] by De Loecker and Warzynski (2012) (we will call the resulting combination the DLW estimator). The approach in (1) is agnostic about consumer demand and firm conduct, and is therefore compatible with a large class of models with imperfect competition.

However, identifying the flexible input elasticity is notoriously difficult. Both traditional and modern estimators suffer from the classic *transmission bias:* input choice is correlated with firm productivity.[4] The envelope condition underlying (1) implies firms endogenously choose flexible inputs as a function of their productivity. When the flexible input expenditure is high, we cannot disentangle whether it is because its output elasticity is high, or because productivity is high.

Gandhi, Navarro, and Rivers (2017) (GNR) shows that the flexible input elasticity is not identified under the standard restrictions on firm input decisions in the proxy variable literature. The empirical strategy proposed in GNR applies the relationship in (1) in the opposite direction – under an assumption about the markup, the flexible input elasticity can be identified using equation (1). This information on the elasticity resolves the source of non-identification in the proxy variable framework, and GNR shows that the remaining structure is sufficient to identify the production function. Unfortunately, this approach strips (1) of any empirical content for estimating markups.

We show that inference about markups through the relation (1) is possible if we

[2] An input is flexible if it is both variable and static, where variable means the input can be variably adjusted to produce today's output, and static means today's input affects today's profits only (and hence has no dynamic consequences).

[3] By far the most popular estimators in this vein are Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg et al. (2015). As of December 2018, these papers collectively have over 11,000 Google Scholar citations.

[4] Marschak and Andrews (1944) first articulated this problem in a general setting. Hoch (1958) developed it further for Cobb-Douglas production functions, coining the "transmit" terminology.

impose more economic structure on the production function. The main structure we consider is that the production function is homogeneous, characterized by a known returns to scale parameter. Constant returns to scale (CRS) is a natural choice in practice. Such a restriction is already common in applied work with firm-level data, as well as macroeconomic models with heterogeneous firms[5]. With a returns to scale parameter, we resolve the empirical problem of separating markups from the flexible input elasticity, allowing for point identification of firm-level markups. Our identification argument gives rise to a natural extension of the standard proxy variable GMM estimators, which makes the returns to scale structure simple to implement in practice. For illustration, we present the moment conditions explicitly for the Cobb-Douglas and translog production functions.

Using a Monte Carlo simulation, we show that the bias from ignoring the flexible input identification problem is large. Non-identification produces biases as large as the levels of markups themselves. In other words, the DLW method can produce markups of 2 when the true markup is 1. In contrast, assuming constant returns to scale reduces the average bias by two orders of magnitude to become economically insignificant, even when the true scale elasticity is not exactly 1. We also explore an alternative source of identification suggested in the literature, serially correlated input prices that vary across firms (DLW). Such variation overcomes the flexible input problem when: (1) firm-level prices are observable; and (2) the input price variation is orthogonal to productivity and output prices. However, these conditions are largely impractical — the former is rarely available in production data, while the latter means the price variation cannot come from quality differences. Firms would need to pay higher prices for inputs because they are unlucky, not because they are purchasing better inputs or because of changes in demand or conduct conditions. Our Monte Carlo evidence demonstrates the insufficiency of lagged inputs as instruments for markup estimation in this setting. Even when these prices are completely orthogonal to productivity, the DLW estimator bias is still roughly one order of magnitude larger than our CRS estimator when the prices are unobserved.

We apply our approach to estimate production functions for US public firms, documenting significant advantages of our estimator on real data. We start by showing that the Basu and Fernald (1997)-implied scale elasticities are about 1, a fact which is not meaningfully changing over time. We then compare our CRS estimator with the traditional DLW estimator. We find that the DLW estimator produces markups

[5] For example, Basu and Fernald (1997), Syverson (2004), and Foster et al. (2008) all find that constant returns to scale is a good approximation to reality.

that are similar to the theoretical upper bound on markups implied by the proxy model alone (without assuming CRS), often even exceeding it. Moreover, these markups are unstable across common specification choices, such as estimating Cobb-Douglas or translog production functions. Under a Cobb-Douglas specification, the DLW estimator produces aggregate market power estimates just above 1; under a translog specification, that statistic jumps above 6.

We examine the macroeconomic implications of non-identification by appealing to the recent literature on the subject.[6] An efficient allocation of resources in standard models implies low and relatively equalized markups – high markup firms should see increased resource flows, decreasing their markups. To quantify the efficiency implications of our results, we rely on the sufficient statistics proposed in Peters (2018). That paper builds a model of innovation, firm dynamics, and heterogeneous markups that measures efficiency losses through lower creative destruction. Fitting our estimates to the model, we find annual output losses of about 1% (100B to 200B USD annually). Further, the total payments to labor are 11 percent lower than an idealized benchmark. These estimates are similar in size to the calibrated estimates in Edmond, Midrigan, and Xu (2018), which uses an alternative heterogeneous firm model to find markup dispersion misallocation losses of 1%. While inference from our estimates is economically large, inference from DLW markups would incorrectly double these wedges.

We further explore the robustness of our findings by extending our approach to better account for firm-specific heterogeneity. We use techniques from machine learning to classify firms within industries based on technological similarity, and use these new groups as our estimating samples. Specifically, we check how much our results differ with the k-means approach of Bonhomme et al. (2017) and Bonhomme et al. (2019), and with a regression tree approach.

Section 1 reviews the proxy variable production model and the identification problem of the flexible input elasticity. Sections 2 and 3 present our solution in the Cobb-Douglas and general nonparametric cases. In Section 4, we offer explicit estimation algorithms for the most commonly used production specifications. Sections 5 and 6 evaluate our estimator against the standard DLW estimator in Monte Carlo and Com-

[6] These models have evolved from recent work in the international economics literature that quantifies markup dispersion and misallocation as a response to trade shocks. They do so by specifying richer parametric models of demand and firm competitive behavior. See Epifani and Gancia (2011), Edmond, Midrigan, and Xu (2015), Edmond et al. (2018), and Peters (2018).

pustat data. Section 7 extends our estimator to address heterogeneous technology concerns, and Section 8 concludes.

1 The Model and the Non-Identification Problem

Let q_t be log output, and (v_t, k_t) be the log of flexible inputs and capital in period t . Flexible inputs v_t are variable and static: they contribute to time t's production, but have no effect on the firm's future decisions. Capital k_t is fixed and dynamic: it can not be adjusted in period t, but it affects period t output.[7] Consequently, the firm treats the capital stock k_t as a state variable at time t (chosen before productivity innovations are known to the firm). The data are inputs and output over the panel $t = 1, \ldots T$,

$$y = \{(q_t, v_t, k_t)\}_{t=1}^{T} \tag{2}$$

and the joint distribution of the data y in the underlying population of firms is identified in the data.

The proxy approach to production function estimation relates these data to a model of production that consists of three parts. We label this set of assumptions the proxy structure:

1. Output and inputs in each period are related in the following way:

   $$q_t = f\left(v_t, k_t\right) + a_t + \epsilon_t \tag{3}$$

   where f is the production function characterizing the technology of an industry, a_t is a productivity shock that the firm observes before making its period t input decisions, and ϵ_t is an ex-post shock that is independent of the firm's input decisions.

2. Productivity is a Markov process,

   $$a_t = g\left(a_{t-1}\right) + \eta_t, \tag{4}$$

   where the shocks η_t are uncorrelated with inputs chosen before period t, so:

   $$\mathbb{E}\left[\eta_t | k_t, v_{t-1}, k_{t-1}, \ldots\right] = 0, \tag{5}$$

[7] More generally, k_t can be a vector of inputs each of which is not flexible, i.e. either fixed, dynamic, or both.

where the ellipses represent all other lags of the inputs.

3. Flexible input demand has the form $v_t = v(a_t, k_t)$ where $v(\cdot, k_t)$ is a strictly increasing function of a_t. The model is called the proxy structure because this assumption implies productivity can be proxied by observable inputs, $a_t = v^{-1}(v_t, k_t) = a(v_t, k_t)$.

Gandhi, Navarro, and Rivers (2017) show that this structure is insufficient to identify f. For any proposed production function f, there exists an $\widetilde{f}$ that is observationally equivalent: both f and $\widetilde{f}$ are consistent with the proxy structure and generate the same joint distribution of observables. Estimators using only these restrictions are not consistent.

GNR also shows that if the flexible input elasticity $f_v(v_t, k_t)$ were identified, then the production function f would be identified over the support of the data given the proxy structure; the source of non-identification is the flexible input elasticity. GNR uses the relationship (1) combined with an assumption on the markup to identify the production function — if the markup is known, then the flexible input elasticities can be recovered, and this exactly identifies the remaining parameters of the production function under the proxy structure. For example, with perfect competition in the output market, markups are all 1, and (1) collapses to measuring the flexible input elasticity directly from the data:

$$\text{Output Elasticity of Flexible Input} = \frac{\text{Expenditure on Flexible Input}}{\text{Revenue}}$$

Our goal is instead to use the relationship to learn about markups, so we cannot use this approach. We need another structural assumption to identify the output elasticity of the flexible input, and therefore the markup.

2 Identification with Constant Returns to Scale for Cobb-Douglas Production Functions

We resolve the same identification problem addressed in Gandhi, Navarro, and Rivers (2017) by adding the restriction that we know the returns to scale of the production function. While the identification argument is nonparametric (see Section 3), we will start with a simpler case that offers the intuition behind our argument. To illustrate how our identifying assumption works, we consider identification for Cobb-Douglas production functions with constant returns to scale (CRS).

Intuitively, CRS identifies the production function in the following way:

1. GNR shows that if the flexible input elasticity is known, then the non-flexible input elasticities are identified under the proxy structure; for each flexible input elasticity, there exists a unique production function.
2. We show that if the non-flexible input elasticities are known, then the flexible input elasticity is identified under CRS; for each set of non-flexible input elasticities, there exists a unique production function.
3. Consequently, we have a fixed point problem in f_v:

$$f_v(v,k) = 1 - f_k(v,k;f_v) \tag{6}$$

   To formally establish identification, we need to show that the fixed point problem has a unique solution.

When the production function is Cobb-Douglas, we have the following equations from the proxy structure introduced above,

$$\begin{aligned} p_t + q_t &= \theta_v v_t + \theta_k k_t + a_t + \epsilon_t && (7)\\ a_t &= g(a_{t-1}) + \eta_t && (8)\\ \mathbb{E}[\eta_t | k_t, v_{t-1}, k_{t-1}, \ldots] &= 0 && (9)\\ \mathbb{E}[\epsilon_t | v_t, k_t] &= 0 && (10)\\ a_t &= a(v_t, k_t) && (11) \end{aligned}$$

The GNR argument implies moment conditions generated by (3) and (4) are insufficient to identify θ_v and θ_k. By assumption (5), v_t is correlated with a_t, so we need to instrument v_t to identify θ_v in equation (1). However, lagged inputs have no power as instruments for the flexible input in the proxy model. To see why, write

$$v_t = v(a_t, k_t) = v(g(a(v_{t-1}, k_{t-1})) + \eta_t, k_t). \tag{12}$$

Conditional on (k_t, v_{t-1}, k_{t-1}), the only variation in v_t is through η_t. However, lagged inputs are uncorrelated with η_t conditional on (k_t, v_{t-1}, k_{t-1}) by assumption (3). In other words, we exhaust our variation in (k_t, v_{t-1}, k_{t-1}) to identify $g(\cdot)$ and the capital elasticity. The remaining available variation in flexible inputs necessarily comes from the productivity shock, or else we couldn't use v_t to proxy for a_t in the first place. But η_t must be orthogonal to lagged inputs, or else they would not be valid instruments.

We resolve the identification problem by adding the assumption of constant returns to scale (in fact, any known returns to scale function will do),

Assumption 1. [Constant returns to scale] $\theta_k + \theta_v = 1$.

Constant returns to scale removes one parameter from the problem so the number of parameters is equal to the number of instruments (the only instrument with power is capital). We can write the production equation as

$$\begin{aligned} p_t + q_t &= \left(1-\theta_v\right) k_t + \theta_v v_t + a_t + \epsilon_t && (13) \\ \implies p_t + q_t - k_t &= \theta_v \left(v_t - k_t\right) + g\left(a_{t-1}\right) + \eta_t + \epsilon_t && (14) \\ &= \theta_v \left(v_t - k_t\right) + \widetilde{g}\left(k_{t-1}, v_{t-1}\right) + \eta_t + \epsilon_t && (15) \\ \mathbb{E}\left[p_t + q_t - k_t | k_t, v_{t-1}, k_{t-1}\right] &= \theta_v \left(\mathbb{E}\left[v_t | k_t, v_{t-1}, k_{t-1}\right] - k_t\right) + \widetilde{g}\left(k_{t-1}, v_{t-1}\right), && (16) \end{aligned}$$

where $\widetilde{g}\left(k_{t-1}, v_{t-1}\right) = g\left(a_{t-1}\left(k_{t-1}, v_{t-1}\right)\right)$. The production function is identified if Assumption 2 holds:

Assumption 2. There exists $\left(k_t, k_{t-1}, v_{t-1}\right)$ such that,

$$\frac{\partial}{\partial k_t} \mathbb{E}\left[v_t | k_t, k_{t-1}, v_{t-1}\right] \neq 1. \tag{17}$$

Choose a $\left(k_t, k_{t-1}, v_{t-1}\right)$ that satisfies the above assumption. Then,

$$\frac{\partial}{\partial k_t} \mathbb{E}\left[p_t + q_t - k_t | k_t, v_{t-1}, k_{t-1}\right] = \theta_v \left(\frac{\partial}{\partial k_t} \mathbb{E}\left[v_t | k_t, k_{t-1}, v_{t-1}\right] - 1\right), \tag{18}$$

and θ_v is identified.

Assumption 2 is both sufficient and necessary.

To see why it is necessary, suppose that the assumption was violated so that

$$\frac{\partial}{\partial k_t} \mathbb{E}\left[v_t | k_t, v_{t-1}, k_{t-1}\right] \equiv 1. \tag{19}$$

Then, by the fundamental theorem of calculus,

$$\mathbb{E}\left[v_t | k_t, v_{t-1}, k_{t-1}\right] - \mathbb{E}\left[v_t | \overline{k}, v_{t-1}, k_{t-1}\right] = k_t - \overline{k}, \tag{20}$$

for some fixed value $\overline{k}$. Then,

$$\begin{aligned}\mathbb{E}\left[p_t + q_t - k_t | k_t, v_{t-1}, k_{t-1}\right] &= \theta_v \left(\mathbb{E}\left[v_t | k_t, v_{t-1}, k_{t-1}\right] - k_t\right) + \widetilde{g}\left(k_{t-1}, v_{t-1}\right) && (21)\\ &= \theta_v \left(\mathbb{E}\left[v_t | \overline{k}, v_{t-1}, k_{t-1}\right] - \overline{k}\right) + \widetilde{g}\left(k_{t-1}, v_{t-1}\right), && (22)\end{aligned}$$

Now the right hand side varies only with (v_{t-1}, k_{t-1}) and $\widetilde{g}\left(k_{t-1}, v_{t-1}\right)$, so we cannot separately identify θ_v and $\widetilde{g}$ because they are collinear.

3 Nonparametric Identification with Returns to Scale

While Cobb-Douglas with constant returns to scale makes the identification argument particularly straightforward, it is not crucial for our results. Theorem 1 presents our general, nonparametric identification result. For ease of exposition, we assume a returns to scale of 1. However, all arguments follow for any known returns to scale.

The main identification assumption aside from a known returns to scale is about the distribution of flexible inputs $\{v_t\}$ conditional on current fixed inputs $\{k_t\}$. Intuitively, the assumption requires that variation in k_t provides sufficient variation in v_t, and that the parts of v_t that are not collinear with k_t are correlated with k_t. For example, identification fails if $v_t = k_t + \varepsilon_t$, where ε_t is uncorrelated with k_t.

3.1 Proof of the General Case

Theorem 1. If, for any function $\Delta(v, k)$ such that for almost all (v, k),

$$-\mathbb{E}\left[\frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t} \Delta\left(v', k_t\right) dv' | k_t = k, v_{t-1}, k_{t-1}\right]\right] + \int_{\overline{v}}^{v} \frac{\partial}{\partial k}\Delta\left(v', k\right) dv' + \Delta\left(v, k\right) = 0 \quad (23)$$

it must be the case that $\int |\Delta(v, k)| = 0$, then either:

1. There exists no production function in the set identified by the proxy structure with constant returns to scale.
2. For any two production functions in the identified set, f^1 and f^0,

$$\int_v \int_k |\left(f^1 - f^0\right)(v, k)| dk dv = 0. \quad (24)$$

Proof. Sufficiency:

Suppose that the only function Δ satisfying,

$$-\mathbb{E}\left[\frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t}\Delta\left(v',k_t\right)dv'|k_t=k,v_{t-1},k_{t-1}\right]\right]+\int_{\overline{v}}^{v}\frac{\partial}{\partial k}\Delta\left(v',k\right)dv'+\Delta\left(v,k\right)=0 \tag{25}$$

is $\Delta = 0$. Suppose, by way of contradiction, that there are two production functions in the identified set such that,

$$\int\int|f^1\left(v,k\right)-f^0\left(v,k\right)|dvdk\neq 0. \tag{26}$$

For a given flexible input elasticity, we can recover the production function using the proxy structure:

$$\begin{aligned} f\left(v,k\right)=\mathbb{E}&\left[q_t-\int_{\overline{v}}^{v_t}f_v\left(v',k_t\right)dv'|k_t=k,v_{t-1},k_{t-1}\right]-\\ \mathbb{E}\left[q_t-\int_{\overline{v}}^{v_t}f_v\left(v',k_t\right)dv'|k_t=\overline{k},v_{t-1},k_{t-1}\right]&+\int_{\overline{v}}^{v}f_v\left(v',k\right)dv' \end{aligned} \tag{27}$$

Because both f^1 and f^0 satisfy the proxy structure, we can expressed them as above. A production function that satisfies the proxy structure has constant returns to scale if

$$\begin{aligned} 1=f_k\left(v,k\right)+f_v\left(v,k\right)=\frac{\partial}{\partial k}\mathbb{E}&\left[q_t-\int_{\overline{v}}^{v_t}f_v\left(v',k_t\right)dv'|k_t=k,v_{t-1},k_{t-1}\right]\\ &+\int_{\overline{v}}^{v}f_{vk}\left(v',k\right)dv'+f_v\left(v,k\right) \end{aligned} \tag{28}$$

Both f^1 and f^0 have constant returns to scale so differencing the equations above gives:

$$0=-\frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t}\left(f_v^1-f_v^0\right)\left(v',k_t\right)dv'|k_t=k,v_{t-1},k_{t-1}\right]+\int_{\overline{v}}^{v}\frac{\partial}{\partial k}\left[f_v^1-f_v^0\right]dv'+f_v^1-f_v^0 \tag{29}$$

Define $\Delta\left(v,k\right)=f_v^1-f_v^0$. Then the identification assumption $\Delta=0$ implies $f_v^1=f_v^0$. Because both have constant returns to scale, this implies that either $f_k^1=f_k^0$, or that the two production functions are the same which is a contradiction of the assumption that they are different. There cannot be more than one production function in the identified set given the identification condition is true.

Necessity:

Suppose there exists a $\Delta(v,k) \neq 0$ for all (v,k) such that,

$$-\mathbb{E}\left[\frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t} \Delta(v',k_t)\,dv'|k_t = k, v_{t-1}, k_{t-1}\right]\right] + \int_{\overline{v}}^{v} \frac{\partial}{\partial k}\Delta(v',k)\,dv' + \Delta(v,k) = 0 \tag{30}$$

then, if there exists a production function f in the identified set, the production function has constant returns to scale, or:

$$f_k(v,k)+f_v(v,k) = \frac{\partial}{\partial k}\mathbb{E}\left[q_t - \int_{\overline{v}}^{v_t} f_v(v',k_t)\,dv'|k_t = k, v_{t-1}, k_{t-1}\right]+\int_{\overline{v}}^{v} f_{vk}(v',k)\,dv' \\ + f_v(v,k) = 1. \tag{31}$$

We can construct another production function $\widetilde{f}$ that satisfies the identifying assumptions in the following way:

1. Let $\widetilde{f}_v = f_v + \Delta$.
2. Applying the same proxy transformation,

$$\widetilde{f} = \int_{\overline{k}}^{k} \frac{\partial}{\partial k}\mathbb{E}\left[q_t - \int_{\overline{v}}^{v_t} f_v(v',k_t)\,dv'|k_t = k, v_{t-1}, k_{t-1}\right]dk + \int_{\overline{v}}^{v} f_v(v',k)\,dv' \\ - \int_{\overline{k}}^{k} \frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t} \Delta(v',k_t)\,dv'|k_t = k, v_{t-1}, k_{t-1}\right]dk + \int_{\overline{v}}^{v} \Delta(v,k) \tag{32}$$

$$\Longrightarrow \widetilde{f} = f - \int_{\overline{k}}^{k} \frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t} \Delta(v',k_t)\,dv'|k_t = k, v_{t-1}, k_{t-1}\right]dk + \int_{\overline{v}}^{v} \Delta(v,k) \tag{33}$$

3. Differentiating gives

$$\widetilde{f}_k = f_k - \frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t} \Delta\left(v', k_t\right) dv' | k_t = k, v_{t-1}, k_{t-1}\right] + \int_{\overline{v}}^{v} \frac{\partial}{\partial k}\Delta\left(v', k\right) dv' \tag{34}$$

$$\widetilde{f}_v = f_v + \Delta\left(v, k\right) \tag{35}$$

$$\widetilde{f}_v + \widetilde{f}_k = 1 - \frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t} \Delta\left(v', k_t\right) dv' | k_t = k, v_{t-1}, k_{t-1}\right] \tag{36}$$

$$+ \int_{\overline{v}}^{v} \frac{\partial}{\partial k}\Delta\left(v', k\right) dv' + \Delta\left(v, k\right)$$

4. Because this must be true for all (v, k, v_{t-1}, k_{t-1}) and the left hand side is only a function of (v, k), integrating both sides with respect to (v_{t-1}, k_{t-1}) makes no difference so:

$$\begin{aligned}\widetilde{f}_v + \widetilde{f}_k = 1 - \mathbb{E}\left[\frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t} \Delta\left(v', k_t\right) dv' | k_t = k, v_{t-1}, k_{t-1}\right]\right] \\ + \int_{\overline{v}}^{v} \frac{\partial}{\partial k}\Delta\left(v', k\right) dv' + \Delta\left(v, k\right) \\ = 1\end{aligned} \tag{37}$$

Implying $\widetilde{f}$ has constant returns to scale. Hence, there are multiple production functions in the identified set, a contradiction.

□

3.2 Interpretation and Application

We can use Theorem 1 to generate identification conditions which are easy-to-check (and empirically testable) for particular functional form choices for the production function. To do so, we can replace Δ with the functional form of f_v to recover what the identification assumption means for that functional form.

The following result is useful for understanding what the identification condition means in the context of particular functional forms:

Result 1. If $\Delta(v,k)$ satisfies,

$$-\mathbb{E}\left[\frac{\partial}{\partial k}\mathbb{E}\left[\int_{\overline{v}}^{v_t}\Delta(v',k_t)\,dv'|k_t=k,v_{t-1},k_{t-1}\right]\right]+\int_{\overline{v}}^{v}\frac{\partial}{\partial k}\Delta(v',k)\,dv'+\Delta(v,k)=0, \tag{38}$$

then it must be the case that Δ is homogenous of degree zero (regardless of data generating process), i.e.

$$\Delta_v+\Delta_k=0 \tag{39}$$

Proof. Differentiate the identification condition with respect to v. □

We present more explicit identifying conditions derived from the general condition above for two common functional form choices: log additively separable production functions (of which Cobb-Douglas is a special case) and translog production functions. Both make use of Result 1.

3.3 Log Additively-Separable Example

Log-additively separable production functions take the form $f=f^1(v)+f^2(k)$. Because Δ is homogenous of degree zero and because $\Delta_k=0$ (f_v does not vary with k), it must be the case that $\Delta_v=0$. So $\Delta(v,k)=\delta$, a constant. The identification condition is then that there does not exist $\delta\neq 0$ such that for all (v,k),

$$-\delta\mathbb{E}\left[\frac{\partial}{\partial k_t}\mathbb{E}[v_t|k_t=k,v_{t-1},k_{t-1}]\right]+\delta=\delta\times\left\{1-\mathbb{E}\left[\frac{\partial}{\partial k_t}\mathbb{E}[v_t|k_t=k,v_{t-1},k_{t-1}]\right]\right\}=0, \tag{40}$$

So as long as there exists some k such that

$$\mathbb{E}\left[\frac{\partial}{\partial k_t}\mathbb{E}[v_t|k_t=k,v_{t-1},k_{t-1}]\right]\neq 1, \tag{41}$$

the production function is identified.

3.4 Translog Example

In the translog functional form, $\Delta(v,k) = \delta_0 + \delta_v v + \delta_k k$. The identification condition is that, there does not exist $\delta \neq 0$ such that

$$\delta_0 \times \left\{ -\mathbb{E}\left[\frac{\partial}{\partial k} \mathbb{E}\left[v | k_t = k, v_{t-1}, k_{t-1} \right] \right] + 1 \right\} \tag{42}$$

$$+\delta_v \times \left\{ -\frac{1}{2}\mathbb{E}\left[\frac{\partial}{\partial k} \mathbb{E}\left[v_t^2 | k_t = k, v_{t-1}, k_{t-1} \right] \right] + v \right\} \tag{43}$$

$$+\delta_k \times \left\{ -\mathbb{E}\left[\frac{\partial}{\partial k} \left(k\mathbb{E}\left[v - \overline{v} | k_t = k, v_{t-1}, k_{t-1} \right] \right) \right] + (v - \overline{v}) + k \right\} = 0 \tag{44}$$

$$\delta_0 w_0 + \delta_v w_v + \delta_k w_k = 0 \tag{45}$$

If there exist at least three (v,k) such that the three vectors (w_0, w_v, w_k) are linearly independent, then the only δ that satisfies the equation is $\delta = 0$. A sufficient condition is:

$$\mathbb{E}\left[w(v,k)\, w(v,k)^\top \right] \delta = 0 \implies \delta = 0 \quad \text{if } \mathbb{E}\left[w(v,k)\, w(v,k)^\top \right] \text{ is invertible.}$$

4 Estimators for Common Functional Forms

In estimation, researchers can use standard proxy estimators with the additional constraint that the production function has constant returns to scale. For simple production functions, it is straightforward to derive closed form conditions. For more complex production functions, we can equivalently add the constant returns to scale restriction as an additional moment restriction in the generalized method of moments problem. Linear-in-parameters functional forms such as Cobb-Douglas have particularly straightforward estimators. In this section, we describe how to apply our method to the multi-step estimation method of Ackerberg et al. (2015) used in De Loecker and Warzynski (2012).

4.1 Simple Production Functions

The Cobb-Douglas production with constant returns to scale can be written as,

$$q_t = \theta_v v_t + \theta_k k_t + a_t + \epsilon_t = v_t + \theta_k (k_t - v_t) + a_t + \epsilon_t \tag{46}$$

$$\implies q_t - v_t = \theta_k (k_t - v_t) + g(a_{t-1}) + a_t + \epsilon_t. \tag{47}$$

Estimation uses the standard proxy estimator, replacing log output with $q_t - v_t$ and replacing the inputs with $k_t - v_t$ and using the instrument k_t.

The translog production function can have constant returns to scale imposed by the parameter restrictions:

$$\theta_k + \theta_{vk} v + 2\theta_{kk} k + \theta_v + \theta_{vk} k + 2\theta_{vv} v = 1 \tag{48}$$

$$\implies \theta_k + \theta_v = 1, \theta_{vk} + 2\theta_{vv} = 0, \theta_{vk} + 2\theta_{kk} = 0 \tag{49}$$

In both of these cases, we can estimate the production function (and hence markups) by simply imposing these parameter restrictions and performing the standard Ackerberg, Caves, and Frazer (2015) steps to estimate the production function. The translog production function has 3 parameters after imposing CRS and the Cobb-Douglas production function has only 1 parameter.

1. Impose the CRS parameter restrictions on the production function $f(v_t, k_t; \theta)$.
2. Regress q_t on a specified transformation of (v_t, k_t) to estimate $\phi(v_t, k_t) = \mathbb{E}[q_t | v_t, k_t]$:

$$q_t = \phi(v_t, k_t) + \epsilon_t \tag{50}$$

3. For a given guess of θ, write $a_t = \phi(v_t, k_t) - f(v_t, k_t)$.
4. Regress a_t on a specified transformation of a_{t-1} to estimate $g(a_{t-1}) = \mathbb{E}[a_t | a_{t-1}]$:

$$a_t = g(a_{t-1}) + \eta_t \tag{51}$$

5. Solve the moment condition for θ:

$$\frac{1}{n} \sum_{it} k_{it} \eta_{it}(\theta) = 0 \tag{52}$$

4.2 Flexible Production Functions

Suppose the production function can be written as,

$$f = r(v, k)^{\top}\theta. \tag{53}$$

For a known vector of functions $r(v, k)$. The model can be estimated in the following way:

1. Regress q_t on some flexible transformation of (v_t, k_t) to estimate $\phi(v_t, k_t) = \mathbb{E}[q_t|v_t, k_t]$.
2. For a given guess of θ,

$$\phi(v_t, k_t) - r(v_t, k_t)^{\top}\theta = a_t \tag{54}$$

3. Regress a_t on (transformations of) a_{t-1},

$$a_t = g(a_{t-1}) + \eta_t \tag{55}$$

4. Let $h(k)$ be a sufficient number of linearly-indepedent transformations of k. Solve the moment equation using GMM,

$$\frac{1}{n}\sum_{it} h(k_{it})\eta_{it}(\theta) = 0$$

$$\frac{1}{n}\sum_{it} r(v_{it}, k_{it})(RTS_{it}(\theta) - 1) = 0 \tag{56}$$

5 Sizing the Problem: Monte Carlo Evidence

In this section, we use a Monte Carlo experiment to demonstrate what can go wrong if we ignore the identification problem and measure markups using lagged flexible inputs as instruments. By comparing our estimator to the De Loecker and Warzynski (2012) estimator, we show that inference on the pattern of markups in an industry can be quite misleading if we do not deal with the identification problem. We also show that reasonable deviations from constant returns to scale (where the true returns to scale is not 1 but we assume it is) have less effect on inference than ignoring the identification problem entirely. If we do not believe that the returns to scale are far from 1, these results suggest that small deviations from constant returns to scale will not overly affect inference.

The data generating process we use for the Monte Carlo experiment assumes each firm solves the following dynamic programming problem,

$$M(A,K) = \max_{V,X,Q_E} Q_E^{\eta+1}\mathbb{E}\left[\exp\left((\eta+1)\epsilon\right)\right] - W_X X - W_V V + \beta\mathbb{E}\left[M\left(A\nu, \delta\times K + X\right)\right] \tag{57}$$

$$\text{st:}\quad A\times V^{\theta}K^{\gamma} \geq Q_E, \tag{58}$$

where the expectation is over ν, a shock to productivity. The dynamic programming problem can be solved analytically, so it is straightforward to draw data from the data generating process.

We parameterize our experiment at the following parameter values (for a given θ and γ),[8]

$$Q = Q_E\times\exp(\epsilon), \quad Q = V^{\theta}K^{\gamma}A, \quad \log\nu\sim N(0,0.25) \tag{59}$$

$$\eta = -0.1, \quad \beta = 0.95, \quad \delta = 0.9, \quad W_X = 1, \quad W_V = 0.4 \tag{60}$$

$$\epsilon\sim N(0,0.5). \tag{61}$$

At these parameter values, the true markup is always $1\frac{1}{9}$. We draw a population of 2000 firms and simulate their actions for 6 time periods, where each firm's initial capital and productivitycomes from the following distribution:

$$\log(K_0-1)\sim N(0,1) \tag{62}$$

$$\log A_0\sim N(0,1). \tag{63}$$

We estimate the model setting $t=6$ (using only data from period $t=5$ and $t=6$). We run the model for a few periods before estimating it to make sure the $t-1$ choice can really be said to have been "chosen" according to the model and not just be a draw from an initial condition.

Our goal is to compare our CRS estimator against the DLW estimator, which is currently the literature's standard for markup estimation. We study how well each

[8] For full details on the solution to the dynamic programming problem, see Appendix A.

method does at estimating average markups and the standard deviation of markups. Table 1 reports the bias of each method for a sample size of 2000 firms simulated from the above model and 10000 Monte Carlo simulations for different values of (θ, γ), some of which satisfy $\theta + \gamma = 1$ and some of which do not.

To be precise, we use the same proxy assumptions for both the DLW estimator and our proposed estimator based on constant returns to scale,

$$q_{it} = \theta v_{it} + \gamma k_{it} + a_{it} + \epsilon_{it} \tag{64}$$

$$a_{it} = \rho a_{it-1} + \log \nu_{it}. \tag{65}$$

What is different between the two models are the moments used in estimation. The identifying moments for both models are,

$$\textbf{DLW Moments:} \quad \mathbb{E}\left[\begin{pmatrix} k_{it} \\ v_{it-2} \end{pmatrix} \log \nu_{it}\right] = 0 \tag{66}$$

$$\textbf{FGT Moments:} \quad \mathbb{E}\left[k_{it} \log \nu_{it}\right] = 1 - \theta - \gamma = 0 \tag{67}$$

Table 1 collects the results of our Monte Carlo experiment. Each row represents a different parameterization of the flexible input elasticity θ and the capital elasticity γ. The first set of columns presents the average bias in the level of markups (where the true value is $1\frac{1}{9}$); the second set of columns presents the average bias in the dispersion of markups (where the true value is 0). We measure bias relative to the estimator we would use if we knew the true production function (to avoid measuring as bias the fact that we only observe ex-post output instead of ex-ante output) — that is, we apply the markup formula replacing the flexible input elasticity with the true flexible input elasticity.

We find the bias from ignoring the identification problem is much larger than the bias from being wrong about the returns to scale in the industry. The average markup estimates using DLW overstate markups substantially, ranging from about 1 to 1.5. In contrast, the bias from our FGT estimator is small, ranging from 0 to 0.3. Our estimate of the standard deviation of markups is also slightly better than DLW — even under misspecification (non-CRS) — but recall: the standard deviations are around two very different means so the distribution of markups is much better captured using our approach.

Figure 1 shows the full distribution of average markup estimates for both the DLW and FGT estimators under constant returns to scale. The non-identification of DLW is

Table 1: Monte Carlo Results

		Bias: Average Markups		Bias: Standard Deviation of Markups	
θ	γ	FGT	DLW	FGT	DLW
0.80	0.20	0.005	1.212	< 0.001	0.008
0.75	0.20	0.159	1.498	0.003	0.004
0.85	0.20	0.279	0.963	0.006	0.011

FGT refers to the method proposed in this paper. DLW refers to applying the proxy method using lagged flexible inputs to instrument for current flexible inputs.

visibly clear from the bimodal distribution of average markups where FGT is normally-distributed around the true value.

5.1 Input Price Variation Is Not a (Practical) Solution

De Loecker and Warzynski (2012) (and more recent work by De Loecker and Eeckhout 2017 and Traina (2018)) uses lags of flexible inputs to instrument for current flexible inputs. That paper's argument is that lags of flexible inputs are correlated with the price of the firm's inputs which will shift the distribution of current inputs. And, since lagged inputs were chosen in the past, they are uncorrelated with later shocks to productivity. Unfortunately, the use of lagged inputs and appeal to unobserved input price variation does not overcome the GNR identification problem.

The basic reason instrumental variable methods based on unobserved input price variation (or other unobserved firm heterogeneity) fail is that either:

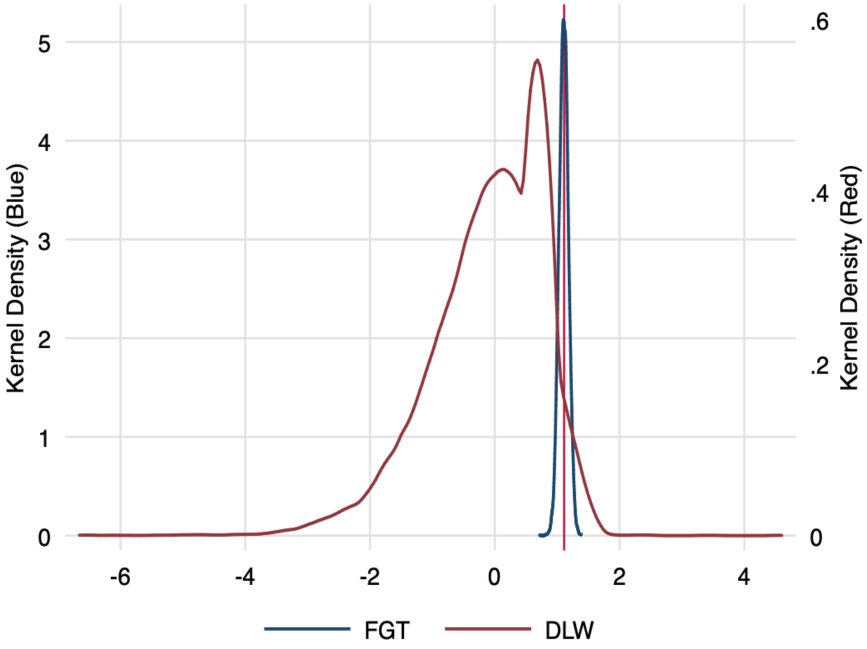
1. There is input price variation so we have an omitted variable (input price) in the residual because input price should be included in the proxy function. So our instrument is correlated with the residual (our instrument is invalid).

2. There is no input-price variation so lagged flexible inputs have no power (our instrument is weak).

In the following section, we establish this point.

If there is no input price variation, then Gandhi et al. (2017) have already shown that

Figure 1: Markup Distribution by Estimator (Vertical Line at True Average)



the model is not identified as we review in Section 1. The problem is fundamental: for lagged inputs to satisfy the exclusion restriction, they must be uncorrelated with η_{it}, but for lagged inputs to have strength as instruments, they must be correlated with η_{it}.

Now consider the case where input prices vary by firm, yielding another firm-specific state variable aside from η_{it} that might shift v_{it}. There remains two general problems: (1) the nature of the input price variation required is more demanding than we might intuit (it is not enough that input prices vary by firm); and (2) the nature of the data required is more demanding than we might see in practice (we must observe the firm-level input prices).

First, suppose input prices vary across firms, and we observe these input prices. In that case, flexible input demand is,

$$v_t = v_t\left(a_t, k_t, w_t\right) = v_t\left(g\left(v_{t-1}, k_{t-1}, w_{t-1}\right) + \eta_t, k_t, w_t\right), \tag{68}$$

where w_t is the log price of flexible inputs.

The first question for identification is whether v_{t-2} has any strength as an instrument; is it correlated with v_t conditional on $(k_t, v_{t-1}, k_{t-1}, w_{t-1})$? Aside from the variables that are conditioned on, there are two state variables that affect flexible input demand: (η_t, w_t). By the proxy structure, v_{t-2} is not correlated with η_t, so for the instrument to have any strength, it must be correlated with w_t, conditional *on* $(k_t, v_{t-1}, k_{t-1}, w_{t-1})$.

To see why this is a stronger condition than we might think, suppose w_t is an AR(1) process,

$$w_{it} = \rho^w w_{it-1} + \text{Innovation}_{it}, \quad \text{Innovation}_{it} \sim N\left(0,1\right). \tag{69}$$

where the innovation is a shock, independent of any decisions the firm makes. Then, conditional on w_{it-1}, the only variation in w_{it} is through the innovation term which is entirely independent of v_{it-2}. If this model is the data generating process, then v_{it-2} has no strength as an instrument. Similarly, if w_{it} were just fluctuations around a firm-specific mean — $w_{it} = \delta_i + \text{fluctuation}_{it}$, where the fluctuations are iid — the model would not be identified.

What we need for v_{it-2} to be an instrument is for there to be a firm-specific component of the wage process.For example,

$$w_{it} = \rho^{w} w_{it-1} + \xi_i + \text{innovation}_{it}. \tag{70}$$

Or that w_t is at least an AR(2) process; we need an underlying state variable that affects w_{it} aside from w_{it-1} for which v_{it-2} can proxy.

The second question for identification is whether v_{t-2} is a valid instrument, i.e. whether its variation is actually exogenous. This question might seem irrelevant because we have already assumed v_{t-2} is uncorrelated with the innovation term in the proxy structure we laid out above. But this structure is justified within a model where inputs are homogeneous and input price variation suggests input quality variation. If input price variation does reflect input quality variation, then innovations in the wage process are related to innovations in the productivity process—when wages go up, it is because the firm is using more productive inputs—in which case v_{t-2} is correlated with η_t.

But the larger issue is that we often do not observe w_t. If input prices vary across firms but we do not observe this variation, then using twice-lagged[9] flexible inputs as instruments will not work. In this case, the proxy function is:

$$v_t = v_t\left(a_t, k_t, w_t\right) \implies a_t = v_t^{-1}\left(v_t, k_t, w_t\right) \tag{71}$$

$$\implies q_t = f\left(v_t, k_t\right) + g\left(v_{t-1}, k_{t-1}, w_{t-1}\right) + \eta_t + \epsilon_t. \tag{72}$$

If we do not observe w_{t-1} and we estimate the above model omitting w_{t-1}, we introduce an omitted variable bias. So the De Loecker and Warzynski (2012) approach requires that we observe wages, which is typically hard to observe for intermediate inputs. In fact, the reason to use lagged flexible inputs as the instrument is that we do not observe firm wages—otherwise, if we believe wage variation is excluded, we

[9] De Loecker and Warzynski (2012) use a value-added production function and treat labor as a flexible input. In that case, labor does not appear in the material demand function, so first-lagged labor is excluded. But when materials are the flexible input and included in the production function, first-lagged materials are not excluded, so we need to use twice-lagged materials (flexible inputs).

Table 2: Monte Carlo Results (Input Price Variation)

		Bias: Average Markups		Bias: Standard Deviation of Markups	
θ	γ	FGT	DLW	FGT	DLW
0.80	0.20	0.251	0.429	0.031	0.053
0.75	0.20	0.119	0.364	0.015	0.047
0.85	0.20	0.176	0.344	0.018	0.035

could just use firm-level wages (w_{it}) directly as the instrument and it would likely be a stronger instrument than indirectly proxying w_{it} via v_{it-2}.

We also compare the two estimators when we introduce firm-level input price variation to the data-generating process. Both estimators are misspecified under this data generating process, but the instruments used in the DLW are no longer weak because lagged inputs are correlated with input price variation; the bias shifts from being a result of the instruments having no power to an omitted variable bias. The data generating process we use for this Monte Carlo experiment is the same as above except that,

$$W_V \sim Uniform(0.3, 0.5),$$

where, previously, we had set $W_V = 0.4$. In Table 2 we present these results. The structure of the table is the same as before, but now with input price variation.

Even under this misspecification which gives lagged flexible inputs power in the first stage (the second stage is still misspecified), our estimator performs better. Intuitively, this is because as the variance of input prices becomes small, our estimator is a better and better approximation to the truth while the DLW estimator is not because it is not identified. There is a set of assumptions under which our estimator is identified and it appears fairly robust to small deviations from those assumptions.

6 Sizing the Problem: US Public Firm Markups

We apply our approach to the Fundamental Annual Compustat file from Wharton Research Data Services. These data span from 1951 to 2017 and cover private sector firms with public equity or debt. Compustat contains firm-level balance sheet information, specifically information on: sales; operating expenditures (OPEX); cost of goods sold (COGS); selling, general, and administrative expenses (SGA); capital; and industry classification.

To select domestic firms, we use standard industry format observations in USD with Foreign Incorporation Codes (FIC) in the USA. For data quality, we include only observations with positive assets, sales, operating expenditures, and gross plants, property, and equipment (PPE). To avoid picking up merger and acquisition distortions, we exclude observations in which acquisitions are larger than 5% of the value of total assets.

Compustat includes only the current SIC code and historical SIC codes starting in 1987. We use historical SIC codes when available. We backfill the industry classification with the first historical SIC code, and replace any remaining missing observations with the current SIC code. To better balance the number of observations in each industry which will be useful for later estimation, we map these SIC codes to the Fama-French 49 industry groups. This mapping is standard in the finance and accounting literature, and roughly combines similar industries that have few observations, and separates industries that have many observations into subindustries. Finally, we exclude utilities (Fama-French 49 code 31) because they are heavily regulated on prices, and financials (Fama-French 49 codes 45 to 49) because their balance sheets are dramatically different from other firms.

We also download price deflators to convert nominal variables to real variables. We use the NIPA Table 1.1.9. GDP deflator (line 1) and nonresidential fixed investment good deflator (line 9).

Just under 9% of sample observations are missing observations of sales, operating expenditures, gross PPE, or net PPE, within a given firm. We replace these missing observations with a linear interpolation of their neighboring values. To get a real measure of sales and flexible inputs, we deflate sales, OPEX, COGS, and SGA by the GDP deflator.

As is standard in the production function estimation literature, we construct our measure of capital using the perpetual inventory method. Specifically, we initialize the capital stock using the first available entry of gross PPE. We then iterate forward on capital using the accumulation equation $k_{it} = k_{it-1} + i_{it} - \delta k_{it-1}$, where we compute net investment using changes to net PPE. Since we want a measure of the real capital stock, we deflate net investment by the investment goods deflator.

Our goal in this paper is not to determine the appropriate specification for production technologyusing Compustat data. For more information on this debate, see Traina (2018) and Diez et al. (2018). Rather, we believe it's important to evaluate the impact of non-identification on a variety of applied specifications. We will focus on the simplest specification (Cobb-Douglas technology with flexible OPEX), but also present results for three others when relevant. In total, we vary {Cobb-Douglas; translog} x {OPEX and PPE; COGS, SGA, and PPE}. Our estimator typically increasingly outperforms as we increase the complexity of the specification.

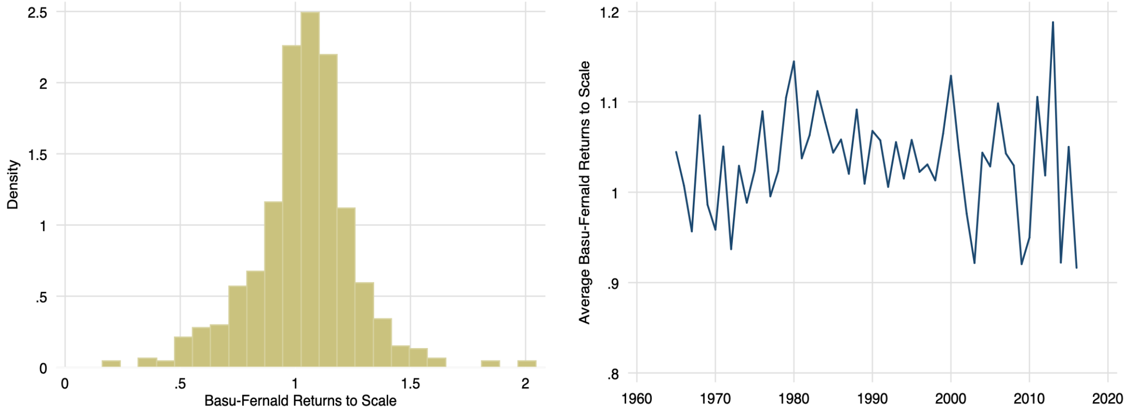
Since we will proceed by imposing constant returns to scale with our benchmark estimator, we first check whether this assumption is reasonable in this particular data setting. To do so, we follow Basu and Fernald (1997), which recommends the following procedure:

1. Generate the cost share of total costs for each input
2. Generate a composite input growth index that sums each input's growth by its lagged cost share
3. Regress output growth on this composite input growth; the coefficient is an estimate for the returns to scale

Although this algorithm does not identify the true returns to scale under typical conditions, it does offer an approximation that has been used in practice (e.g. Syverson (2004)). We follow the procedure for each industry in each year to produce a distribution of returns to scale estimates. Figure 2 collects the results. The left panel presents the distribution of scale elasticity estimates, and the right panel shows the time series of their cross-sectional averages.

The left panel shows that the typical industry-year exhibits constant or slightly increasing returns to scale, with a mean of 1.04. There is some variation away from our CRS benchmark, with a standard deviation of 0.22, which likely represents a combination of true differences in scale elasticities, sampling error, and misspecification. The right panel shows that there is no discernable time trend in these estimates. Regressing our scale elasticity estimates on a linear time trend returns a coefficient that is statistically indistinguishable from zero.

Figure 2: Basu-Fernald Returns to Scale Estimates



6.1 Evaluation with Partial Identification Bounds

While the proxy structure does not point identify the production function in the presence of flexible inputs without further assumption, Flynn (2015) and Flynn and Gandhi (2018) show that it does partially identify the production function by offering lower and upper bounds on the flexible input elasticity. In this section, we use these bounds as evaluation criteria for our estimator against the DLW estimator.

The proxy structure's upper bound is simply the markup we would compute if we estimated the production function via OLS, ignoring endogeneity. The proxy structure tells us that flexible input demand is strictly increasing in productivity: $a_t = v^{-1}(v_t, k_t)$. As a consequence, this structure implies that the inverse flexible input demand is increasing as well, so that $\frac{\partial}{\partial v_t}\mathbb{E}[q_t|v_t, k_t] = \frac{\partial f}{\partial v_t} + \frac{\partial a}{\partial v_t} \geq \frac{\partial f}{\partial v}(v_t, k_t)$. This condition offers an upper bound on the flexible input elasticity, and so an upper bound on markups: $\frac{P^Q Q}{P^V V} \times \frac{\partial}{\partial v_t}\mathbb{E}[q_t|v_t, k_t] \geq \frac{P^Q Q}{P^V V} \times \frac{\partial f}{\partial v}(v_t, k_t) = \mu$. Intuitively, ignoring transmission bias by estimating the production function via OLS will return estimates that are biased strictly upwards under the proxy structure.

The markup lower bound is even simpler: under the proxy structure, markups should never fall below 1. Typical theoretical justifications for markups below 1, such as dynamic pricing, are ruled out by the assumption that the relevant input is flexible, and therefore has no effect on future sales. In other words, to measure markups from the firm's first order condition as in Hall (1988), we must assume away mechanisms that cause markups to fall below 1. Our approach is similar in spirit to Hall (2018),

Table 3: Specification Error with Partial Identification Bounds, Raw Percent

Specification		Too Low		Too High	
Form	Flex	FGT	DLW	FGT	DLW
CD	OPEX	0.16	0.04	0.03	0.55
CD	COGS	0.31	0.35	0.41	0.42
TL	OPEX	0.18	0.51	0.62	0.07
TL	COGS	0.41	0.00	0.00	0.45

which evaluates the original Hall (1988) method by interpreting markups below 1 as sampling error.

Table 3 summarizes how the estimators perform relative to these bounds. Each row represents a different production function specification: {Cobb-Douglas; translog} x {OPEX and PPE; COGS, SGA, and PPE}. The "Too Low" columns report the share of markups below the lower bound of 1 under each estimator, whereas the "Too High" columns report the share above the upper bound. We aggregate these shares on a cost-weighted basis, reasoning that the flexible input cost is the relevant weight for welfare calculations per Edmond et al. (2018).

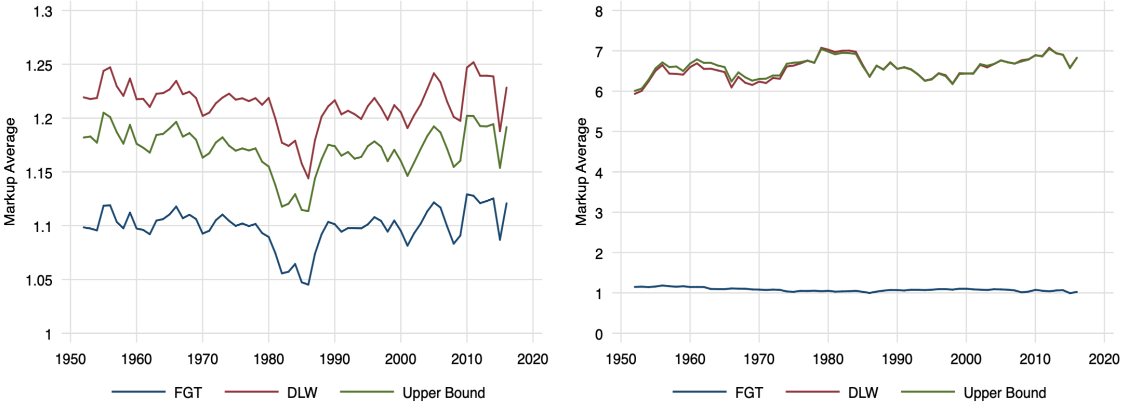
Our estimator typically performs substantially better at not exceeding the upper bound, and performs slightly worse at not falling below the lower bound. However, this takeaway varies across production function specification. While Table 3 shows how often markups fall outside the partial identification bounds, it does not show how substantial these violations are. To check the quantitative importance of these bound violations, we next look at the average bias they induce. We measure each markup estimate's distance past the bound, setting within-bound estimates to zero. Table 4 presents the averages of these bound violations.

While our estimator still has some notable biases, a key feature is that they are not substantial. The largest average bias produced is 0.13, and a typical bias is about 0.05. In contrast, the DLW estimator produces a dispersed set of biases, including some very large ones. To level the playing field away from obvious outliers, the rest of the paper subsets on markups between 0.1 and 10, which we consider to be a priori reasonable bounds for anomalies.

Table 4: Specification Error with Partial Identification Bounds, Average Violation

Specification		Too Low		Too High	
Form	Flex	FGT	DLW	FGT	DLW
CD	OPEX	0.02	0.00	0.00	0.06
CD	COGS	0.04	0.06	0.11	0.02
TL	OPEX	0.02	4.99	0.06	0.23
TL	COGS	0.13	0.00	0.00	0.27

Figure 3: The Level of Markups, by Estimator

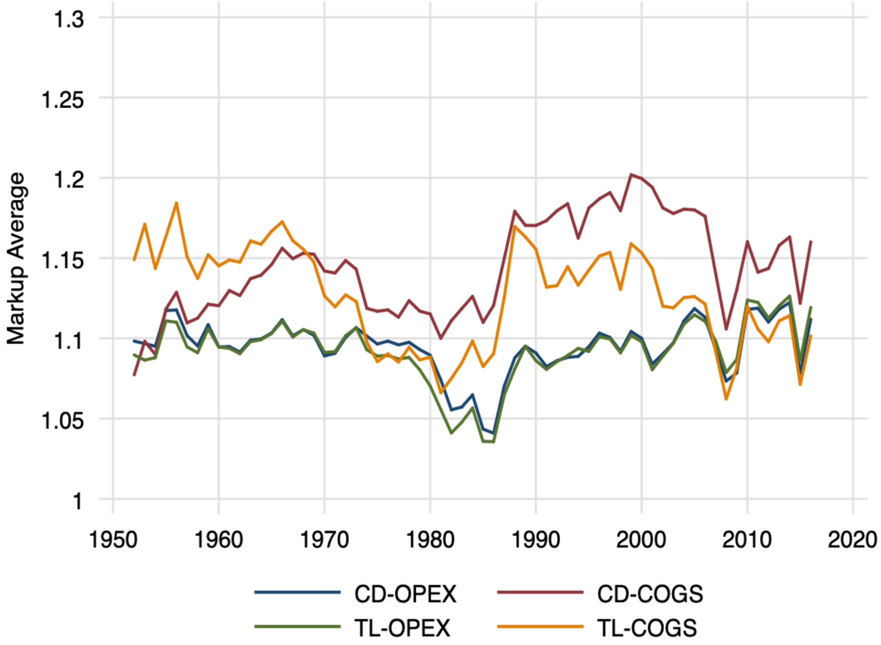


6.2 Stability to Alternative Specifications

Motivated by the partial identification results, we next explore the stability of the estimators by looking at how their resultant markups change across the different production technology specifications. Figure 4 presents the time series of aggregate market power estimates for the four considered in this paper. The left panel is the simplest specification – Cobb-Douglas OPEX as used in Traina (2018). The right panel is the most complex specification – translog COGS-SGA, similar to the robustness section of Diez et al. (2018). The blue line represents our CRS estimator, the red line represents the DLW estimator, and the green line represents the upper bound implied by Flynn (2015) and Flynn and Gandhi (2018). As before, these estimates use flexible input cost-weights to aggregate.

As hinted by the earlier tables, the DLW estimator is typically close to or exceeds

Figure 4: The Stability of the FGT Estimator

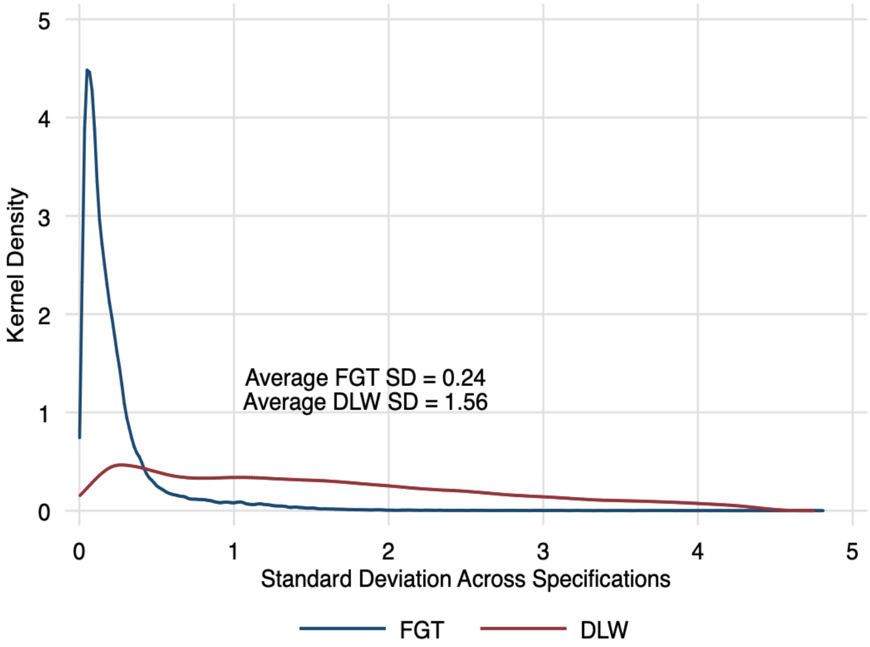


the theoretical upper bound. In the left panel, the red line is about 0.05 higher than the green line, showing that the DLW estimator's upward bias completely swamps any downward bias. More notably for these figures, however, is just how high the theoretical upper bound can reach. In the right panel, both the red and green lines fluctuate between 6 and 7, much higher than in the left panel and more generally implausibly high. The FGT estimator, by comparison, looks about the same across specifications.

Figure 5 confirms this stability by collecting our results across the four specifications. Our estimator is remarkably stable to different specifications of the production function. The differences in aggregate means are low, typically within 0.05 to 0.10.

These estimates are at the aggregate level, however, which is more useful for stylized facts but less useful for firm-level analyses. To quantify how much the markup estimates vary at the micro level, we calculate the standard deviation across the four specifications for each estimator. We therefore have a dispersion estimate for each firm-year observation. Figure 6 presents the distribution of these dispersion estimates

Figure 5: How Much Do Firm-Level Markups Vary by Specification?



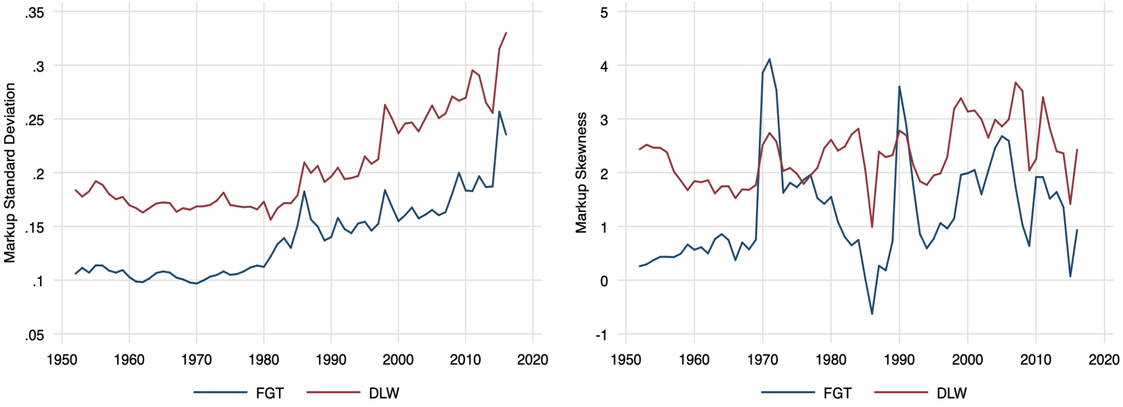
for the two estimators.

Our estimator typically produces markups that are robust to alternative production technology specifications. The average standard deviation is 0.24, against 1.56 for the DLW estimator. We would like to emphasize that these statistics are after truncating the distributions at 0.1 and 10 to remove anomalous outliers, which plague the DLW but not the FGT estimator.

6.3 Differences in Inferences

To give the DLW estimator its best shot, we move forward with the simplest specification (Cobb-Douglas OPEX). Figure 7 shows the differences in markup dispersion by estimator. The left panel measures the cross-sectional standard deviation, whereas the right panel measures the cross-sectional skewness (both weighted by OPEX). In the left panel, the blue line is stable at about 0.10 from 1950 to 1970, when it rises steadily to about 0.20 in 2010. This dispersion occurs around a mean of about 1.15

Figure 6: The Dispersion of Markups, by Estimator



throughout the sample. The rise is notably linear, and there are no otherwise obvious patterns beforehand.

The red line in the left panel is similar in the qualitative pattern, yet at a different y-axis scale. In this case, the blue line starts relatively flat between 0.15 and 0.20 from 1950 to 1980, when it also rises to about 0.30 in 2017. Although the fact that it captures a comparable pattern is reassuring that we're measuring important underlying economic phenomena, the quantification deviates from our identified model in two important ways. First, it meaningfully overstates the level of dispersion, particularly in the earlier parts of the sample by about 50%. Second, because of this overstatement, it significantly underestimates the relative size of the increase by about 50%. The right panel shows the cross-sectional skewness of the two estimators. The FGT estimator is a bit more volatile year-to-year, but typically at a lower level than the DLW estimator. The two estimators do not exhibit any time series trend.

To benchmark how significant these results are, we use the sufficient statistic approach described in Peters (2018) to estimate the corresponding implied output losses from misallocation. The model builds on the canonical Klette and Kortum (2004) framework by adding imperfect competition in product markets to generate variable markups. Given a Pareto distribution of markups, we can approximate the losses to output and the labor share based on the Pareto parameter. Hence to match our estimates to the model, we fit a Pareto distribution of markups each year, and use the model to back out the implied losses.

Figure 7: The Cost of Markups

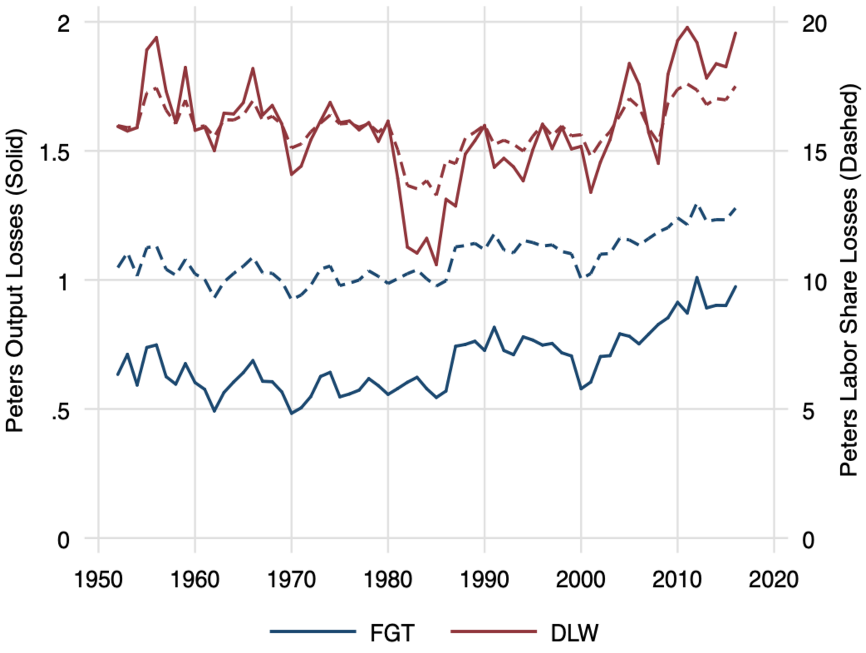


Figure 8 reports our results. Here we see the stark differences in welfare evaluation and macroeconomic implications based on non-identification. Our identified estimator shows output losses hovering just above 0.5% annually, then rising to 1% today. For reference, 1% of US output today is about $200B, about half the budget of the US Department of Defense. While this statistic is large, the DLW estimator implies output losses of roughly twice the scale. The labor share comparison is less stark; we estimate a labor share that's about 11% lower than an idealized benchmark because of markups, whereas the DLW approach estimates about 17%.

7 Robustness Checks with Classification Methods

For robustness, we consider using classification methods from the machine learning literature to allow for more flexible production functions. We suppose that there are groups of firms within each industry with the same production function, but that these groups are unknown to us so we must learn them from the data unlike in the above results where we group the firms ex-ante.

The first method we consider is to group firms (within an industry) with similar input cost shares. Define the share of services in variable expenditures as (SGA/OPEX). Economic theory suggests that firms with similar output elasticities will have similar fractions of their variable spending attributed to each of the variable inputs. We build a regression tree on the logit transformation of (SGA/OPEX) using firm dummies and year dummies as the covariates and leave-one-out least squares cross validation to score the model.

Regression trees work by splitting each set of covariates (firm dummies and year dummies) into binary groups and then estimating the leave-one-out least-squares cross validation metric using those group dummies as covariates. They iteratively split the sample in a hierarchical fashion to maximally reduce the score. They stop splitting when they fail to reduce the cross-validation score. Within each group created by this regression tree, we estimate a different production function using the method we develop in this paper.

The second method we consider is k-means clustering, which allows us to group firms that are similar across a vector of characteristics. Within each industry, we group firms for the vector of (COGS/SALES, SGA/SALES, SALES). We use k=3 to form

Table 5: Specification Error with Partial Identification Bounds

Raw Percent		Too Low			Too High		
Form	Flex	Base	k-Means	Tree	Base	k-Means	Tree
CD	OPEX	0.16	0.10	0.08	0.03	0.17	0.36
CD	COGS	0.31	0.37	0.22	0.41	0.48	0.60
TL	OPEX	0.18	0.10	0.16	0.62	0.65	0.63
TL	COGS	0.41	0.23	0.32	0.00	0.00	0.04

Average Violation		Too Low			Too High		
Form	Flex	Base	k-Means	Tree	Base	k-Means	Tree
CD	OPEX	0.02	0.01	0.01	0.00	0.02	0.04
CD	COGS	0.04	0.04	0.06	0.11	0.13	0.12
TL	OPEX	0.02	0.01	0.07	0.06	0.08	0.09
TL	COGS	0.13	0.05	0.16	0.00	0.00	0.00

three groups within each industry and estimate the production function within each group.

Figure 8: Level and Dispersion Stability

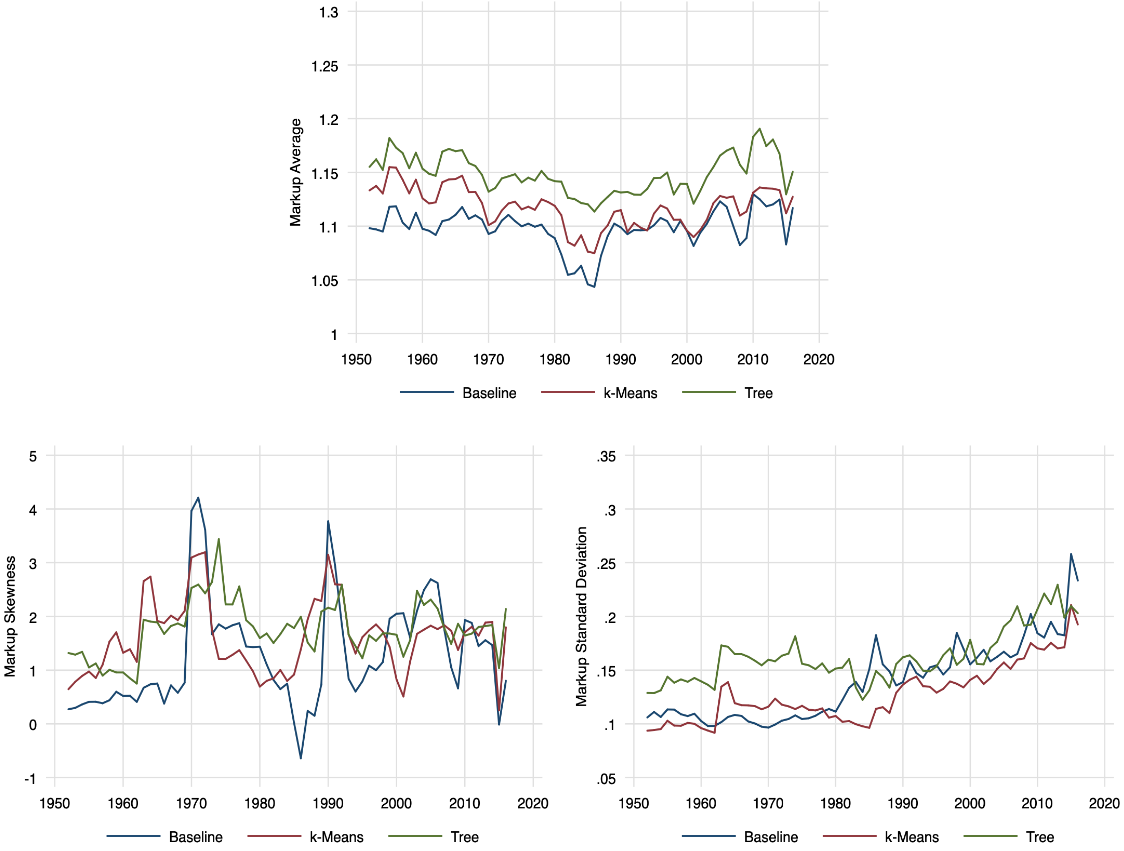
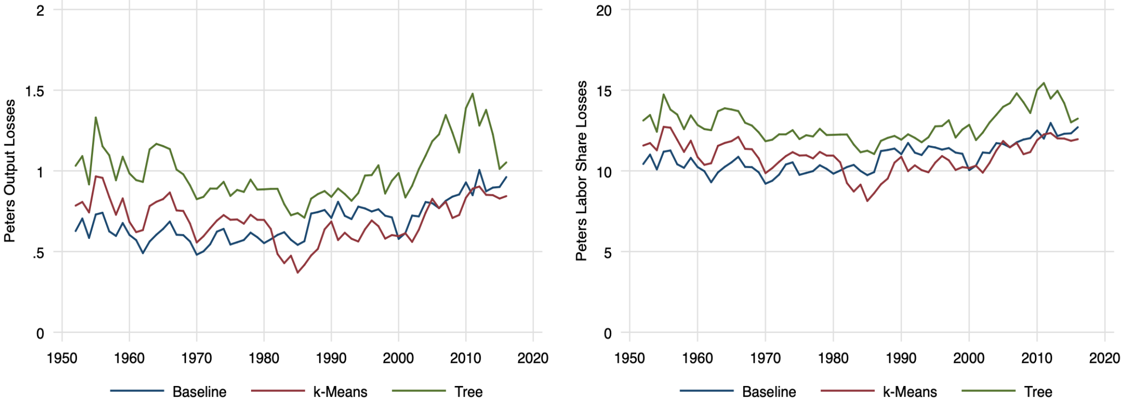


Figure 9: Macroeconomic Implications Stability



8 Concluding Remarks

Recent work in macroeconomics and industrial organization uses the firm's first order condition on flexible inputs to estimate markups from production data. This method offers a direct measure of market power, unlike market concentration or profitability which conflates market power and productivity. Moreover, it does not impose assumptions on demand or market structure, and is therefore consistent with a broad class of economic models.

However, this approach falls victim to the non-identification critique in Gandhi, Navarro, and Rivers (2017) – the parameters in the first order condition are not identified by the most popular production estimators. Worse still, existing solutions to the non-identification problem do not carry over to the setting of market power and imperfect competition. In practice, this problem means that statistical software will return parameters that use variation from pure specification error, making the results non-interpretable.

In this paper, we present a solution that relies on specifying the returns to scale of production. Under such an assumption, we solve the identification problem, and consequently offer applied researchers a practical tool to reliably measure market power in production datasets. We recommend benchmarking results with constant returns to scale, which prior work has found to be a good approximation of reality (Basu and Fernald (1997), Syverson (2004), Foster et al. (2008)). Researchers can then check the robustness of their findings through sensitivity analysis.

We show the empirical success of our solution by evaluating our estimators on both simulated and real world data. In our Monte Carlo experiments, we find the bias from current non-identified approaches is large. Markups typically range from 1 to 2. And current approaches display a bias in our simulated data ranging from 0.34 to 1.50 depending on the data generation process. In contrast, our estimator displays a bias ranging from 0.00 to 0.28. In our application to public firms, we show our estimator has advantages on several dimensions. Most notably, it is robust to technology misspecification, and produces markup estimates that are within the bounds of simple economic reasoning. The difference in macroeconomic implications is stark – our estimator implies output losses from imperfect competition are about half as large as those using current estimators.

We close the paper by showing that remaining problems in production estimators may improve with emerging econometric methods that account for heterogeneous technology, offering a new direction for future research.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–2451. 3, 4, 4.1

Basu, S. and J. G. Fernald (1997). Returns to scale in US production: Estimates and implications. *Journal of Political Economy 105*(2), 249–283. (document), 5, 6, 8

Bonhomme, S., T. Lamadon, and E. Manresa (2017). Discretizing unobserved heterogeneity. (document)

Bonhomme, S., T. Lamadon, and E. Manresa (2019). A distributional framework for matched employer employee data. *Econometrica*. (document)

De Loecker, J. and J. Eeckhout (2017). The rise of market power and the macroeconomic implications. 5.1

De Loecker, J. and F. Warzynski (2012). Markups and firm-level export status. *American Economic Review 102*(6), 2437–71. (document), 4, 5, 5.1, 5.1, 9

Diez, F., D. Leigh, and S. Tambunlertchai (2018). Global market power and its macroeconomic implications. 6, 6.2

Edmond, C., V. Midrigan, and D. Y. Xu (2015). Competition, markups, and the gains from international trade. *American Economic Review 105*(10), 3183–3221. 6

Edmond, C., V. Midrigan, and D. Y. Xu (2018). How costly are markups? (document), 6, 6.1

Epifani, P. and G. Gancia (2011). Trade, markup heterogeneity and misallocations. *Journal of International Economics 83*(1), 1–13. 6

Flynn, Z. (2015). Bounds on statistics of the productivity distribution. *Working Paper*. 6.1, 6.2

Flynn, Z. and A. Gandhi (2018). Partial identification of production functions with flexible inputs. *Working Paper*. 6.1, 6.2

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review 98*(1), 394–425. 5, 8

Gandhi, A., S. Navarro, and D. Rivers (2017). On the identification of production functions: How heterogeneous is productivity? *Conditionally Accepted, Journal of Political Economy*. (document), 1, 2, 5.1, 8

Hall, R. E. (1988). The relation between price and marginal cost in US industry. *Journal of Political Economy 96*(5), 921–947. (document), 6.1

Hall, R. E. (2018). New evidence on the markup of prices over marginal costs and the role of mega-firms in the us economy. 6.1

Hoch, I. (1958). Simultaneous equation bias in the context of the cobb-douglas production function. *Econometrica*, 566–578. 4

Klette, T. J. and S. Kortum (2004). Innovating firms and aggregate innovation. *Journal of Political Economy 112*(5), 986–1018. 6.3

Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies 70*(2), 317–341. 3

Marschak, J. and W. Andrews (1944). Random simulataneous equations and the theory of production. *Econometrica 12*, 143–205. 4

Olley, S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica 64*, 1263–1297. 3

Peters, M. (2018). Heterogeneous mark-ups, growth and endogenous misallocation. (document), 6, 6.3

Syverson, C. (2004). Market structure and productivity: A concrete example. *Journal of Political Economy 112*(6), 1181–1222. 5, 6, 8

Syverson, C. (2019). Macroeconomics and market power: Facts, potential explanations, and open questions. *Brookings Center on Regulation and Markets*. 1

Traina, J. (2018). Is aggregate market power increasing? Production trends using financial statements. *Stigler Center New Working Paper Series No. 17*. 5.1, 6, 6.2

A Solution to the Monte Carlo Dynamic Programming Problem

To solve the dynamic programming problem, write out the first order conditions for (V, X, Q_E),

$$(V): \quad W_V = \lambda \times A\theta V^{\theta-1} K^{\gamma} \tag{73}$$

$$(X): \quad W_X = \beta \mathbb{E}\left[M_K\left(A\eta, \delta K + X\right)\right] \tag{74}$$

$$(Q_E): \quad \left(\eta+1\right) Q_E^{\eta} \mathbb{E}\left[\exp\left(\left(\eta+1\right)\epsilon\right)\right] = \lambda \tag{75}$$

The solution to the static part of the optimization problem is found by collapsing the (Q_E, V) first order conditions,

$$W_V = \left(\eta+1\right)\left(AV^{\theta}K^{\gamma}\right)^{\eta} \mathbb{E}\left[\exp\left(\left(\eta+1\right)\epsilon\right)\right] \times A\theta V^{\theta-1} K^{\gamma} \tag{76}$$

$$\implies V\left(A, K\right) = \left\{\frac{W_V \times A^{-\eta-1} K^{-\gamma(\eta+1)}}{\left(\eta+1\right)\mathbb{E}\left[\exp\left(\left(\eta+1\right)\epsilon\right)\right]}\right\}^{\frac{1}{\theta(\eta+1)-1}} \tag{77}$$

We can solve for the choice of X by using the envelope theorem and the first order conditions to write,

$$M_K\left(A, K\right) = \frac{\gamma\left(\eta+1\right)}{K} \times Q_E^{\eta+1}\left(A, K\right)^{\eta+1} \mathbb{E}\left[\exp\left(\left(\eta+1\right)\epsilon\right)\right] + \beta\delta\mathbb{E}\left[M_K\left(A\eta, \delta \times K + X\right)\right] \tag{78}$$

$$= \frac{\gamma\left(\eta+1\right)}{K} \times Q_E^{\eta+1}\left(A, K\right)^{\eta+1} \mathbb{E}\left[\exp\left(\left(\eta+1\right)\epsilon\right)\right] + \beta\delta W_X \tag{79}$$

Defining $K' = \delta \times K + X$ we can the X first order condition as,

$$\left(1-\beta\delta\right) W_X = \beta\mathbb{E}\left[\exp\left(\left(\eta+1\right)\epsilon\right)\right]\mathbb{E}\left[\frac{\gamma\left(\eta+1\right)}{K'} \times Q_E\left(A\nu, K'\right)^{\eta+1}\right] \tag{80}$$

Because,

$$Q_E\left(A, K\right) = A \times \left\{\frac{W_V \times A^{-\eta-1} K^{-\gamma(\eta+1)}}{\left(\eta+1\right)\mathbb{E}\left[\exp\left(\left(\eta+1\right)\epsilon\right)\right]}\right\}^{\frac{\theta}{\theta(\eta+1)-1}} \times K^{\gamma} = A^{-(\eta+1)\frac{\theta}{\theta(\eta+1)-1}+1} \times \iota_0 K^{\iota_1}, \tag{81}$$

where ι_0 and ι_1 are just functions of the problem's parameters.

Then, the first order condition for X is,

$$(1-\beta\delta)\,W_X = \beta\mathbb{E}\left[\exp\left((\eta+1)\,\epsilon\right)\right]\gamma\,(\eta+1) \times \iota_0 K'^{\iota_1(\eta+1)-1} \times A^{-\frac{(\eta+1)}{\theta(\eta+1)-1}}\,\mathbb{E}\left[\nu^{-\frac{(\eta+1)}{\theta(\eta+1)-1}}\right]. \tag{82}$$

Because ν is log normal (recall that $\mathbb{E}\left[\exp\left(tX\right)\right]$ is the moment generating function of X so we simply evaluate the moment generating function of $\log\nu$ at the relevant value),

$$\mathbb{E}\left[\exp\left(-\frac{(\eta+1)}{\theta\,(\eta+1)-1}\times\log\nu\right)\right] = \exp\left(-\frac{(\eta+1)}{\theta\,(\eta+1)-1}\mu_\nu + \frac{\sigma_\nu^2}{2}\times\left(\frac{(\eta+1)}{\theta\,(\eta+1)-1}\right)^2\right) \tag{83}$$

In any case, we can now solve the first order condition for X in terms of the parameters of the problem which yields the not-so-beautiful but easy-to-solve expression,

$$K'^{\iota_1(\eta+1)-1} = \frac{(1-\beta\delta)\,W_X}{\beta\mathbb{E}\left[\exp\left((\eta+1)\,\epsilon\right)\right]\gamma\,(\eta+1)\,\iota_0\exp\left(-\frac{(\eta+1)}{\theta(\eta+1)-1}\mu_\nu + \frac{\sigma_\nu^2}{2}\times\left(\frac{(\eta+1)}{\theta(\eta+1)-1}\right)^2\right)}A^{\frac{(\eta+1)}{\theta(\eta+1)-1}} \tag{84}$$