Estimating Production Functions with Partially Latent Inputs

Minji Bang, Wayne Gao, Andrew Postlewaite, and Holger Sieg*
University of Pennsylvania

October 5, 2020

Abstract

This paper develops a new method for estimating production functions when the inputs are partially latent. We show that a combination of matching and IV techniques can be used to overcome the problem of partially latent inputs. We propose a corresponding semiparametric estimator, establish its asymptotic distribution, and demonstrate its finite-sample performance in a Monte Carlo study. We then illustrate the usefulness of our approach using two applications. Our first application focuses on the industrial organization of pharmacies. We show that production function differences between chains and independent pharmacies may partially explain the observed transformation of the industry structure. Our second application investigates education production functions and illustrates important differences in child investments between married and divorced couples.

Keywords: production functions, latent variables, endogeneity, semiparametric estimation, instrumental variables, matching.

*We would like to thank Claudia Goldin for comments and suggestions.

1 Introduction

The objective of this paper is to develop and implement a new method for estimating production functions when the input variables are partially latent. We show that a combination of matching and IV techniques can be used to overcome the problem of partially latent inputs. We characterize the asymptotic properties of our estimator and show that it performs well in a Monte Carlo study. We then illustrate the usefulness of our approach using two applications. Our first application focuses on the industrial organization of pharmacies which has undergone a dramatic change during the past two decades: an industry that used to be primarily dominated by local independent pharmacies has been transformed by the entry of large chains that operate in multiple markets. It is important to understand whether this transformation has been driven by technological change.[1] We show that pharmacy chains have different production functions than independent pharmacies, which may partially explain the observed transformation of the industry structure. The second application investigates education production functions. Here we find that there are important differences in child investments between married and divorced couples.

The starting point of our analysis is the canonical model that underlies most of the literature of production function estimation. Recall that firms operate in competitive labor markets.[2] Each firm is subject to a random productivity shock that is uncorrelated with local labor market conditions. We assume each firm's input choices do not necessarily have to be optimal, but are monotonically increasing in the productivity shock. Hence, there is a standard transmission bias problem since inputs are correlated with unobserved productivity shocks (Marschak and Andrews, 1944). In the absence of panel data, researchers typically rely on instrumental variables to overcome this problem.[3] In particular, the key assumption is that differences in local input prices give rise to differences in input choices that are uncorrelated with productivity shocks at the local level, i.e., local input prices can serve as valid instruments for endogenous input choices.[4]

[1]See Goldin and Katz (2016) for a detailed description of this transformation.

[2]For a survey, see Griliches and Mairesse (1998) and Ackerberg et al. (2007).

[3]See Griliches and Mairesse (1998) for a critical discussion of the assumption that these input prices are exogenous. For example, the observed price differences across firms cannot capture input quality differences across firms or different firms' choices of location on a downward sloping input supply curve.

[4]Of course, this is only one approach for resolving the transmission bias problem. Another approach to address this endogeneity problem involves using panel data with fixed effects, first advocated by Hoch (1955, 1962) and Mundlak (1961, 1963). These approaches to identification have been combined with timing assumptions to construct a control function estimator as discussed for example in Olley and Pakes (1996) and Blundell and Bond (1998, 2000), Levinsohn and Petrin (2003), and Ackerberg, Caves, and Frazer (2015).

In this paper, we consider the case in which the econometrician faces a latent variable problem in estimation. This problem arises when the data is collected at the individual worker-level and only contains information for the specific surveyed worker and a limited amount of firm-level information, but no individual information about other workers in the same firm. These types of unstructured data sets are becoming increasingly more prevalent in empirical work. We focus on the case in which each firm needs to hire two types of workers: managers and regular employees.[5] The key problem that arises in estimation is that we observe *either* output and managerial input *or* output and employee input, i.e. we never observe the output, managerial inputs, and employee inputs at the same time. This problem is similar to the latent variable problem encountered in the program evaluation literature. Rubin (1973, 1974) advocates the use of matching in the absence of randomized controlled trials. In the spirit of this approach, we suggest a similar solution to our problem: matching managers and employees that work for similar firms within the same labor market.[6]

Matching ideally should be done based on the unobserved firm-specific productivity shock. That is not feasible. Instead, we match based on the observed output level, which is however only measured with error. If there were no measurement error in output, our assumptions would imply that output is monotonically increasing in productivity, holding local wages constant, and thus there is a one-to-one mapping between the unobserved productivity shock and the observed output level within each local market. This insight then suggests that we can use a matching algorithm to impute the missing input choice. In particular, we can impute the missing input by matching a firm, for which we observe the output and the hours of the manager, with a firm in the same market, for which we observe the same output level and the hours of the employee. In practice, since output is observed with measurement error, it is not guaranteed to be monotone in productivity. However, conditionally expected output, with the measurement error averaged out, remains monotone. We then show that matching can be done based on the conditional expectation of output given the observed labor inputs and wages in the market.

In finite samples, we first nonparametrically estimate the conditional expectation of output, and then use it to match firms and impute latent inputs. Finally we estimate the parameters

We discuss the extension of our methods to this scenario in the conclusions.

[5]We also show how to extend this approach to account for firms with more than one manager and one employee. It is also straightforward to generalize our approach to account for capital or other inputs such as intermediate goods.

[6]In the case of education production functions, we often only measure the inputs of one of the two parents if the parents are divorced.

of the production function using an IV estimator based on the imputed inputs. We provide both high-level and lower-level conditions under which this semiparametric two-step estimator is consistent and asymptotically normal at the usual parametric rate of convergence. We also show that we can obtain efficiency gains by using the estimated conditional expectations of output instead of the observed output in the second-stage IV regression. The technical proof is based on the general econometric theory on semiparametric two-step estimation as in Newey (1994), Newey and McFadden (1994) and Chen, Linton, and Van Keilegom (2003).

To evaluate the performance of our estimator we conduct a variety of Monte Carlo experiments. Our findings suggest that our estimator is well-behaved in samples that are similar in size to those observed in our applications discussed below. We also study the behavior of our estimator when we pool observations across markets as is often necessary for many practical applications. Moreover, we consider other realistic deviations such as the case in which wages are also partially latent.

We then illustrate the usefulness of our new technique using two new applications. First, we apply our new estimator to study differences in productivity in an important industry: pharmacies. Goldin and Katz (2016) have forcefully argued that this is one of the most egalitarian and family-friendly professions in which females face little discrimination in the workforce. One potential explanation of this fact has been related to the rise of chains which have replaced independent pharmacies in many local markets. It is, therefore, useful to test the hypothesis whether chains have access to better production technologies than independent pharmacies.

We use data from that National Pharmacist Workforce Survey in 2000. One of the key advantages of this survey is that it not only collects data for each pharmacist that is surveyed but also a limited amount of information at the store level. In particular, we know how many pharmacists work in the store, and we observe the store's output measured by total prescriptions. The data suggest that there are some potentially important differences between independent and chain pharmacies. For example, chains operate longer hours and, therefore, managers and employees work longer hours in chains than in independent pharmacies.

We then implement our new estimator to measure differences across firm types. We restrict the sample to pharmacies with no more than 3 pharmacists, i.e., pharmacies with one manager and up to two employees who are also trained as pharmacists. We show that this is not a restrictive assumption in our sample. The output is measured by prescriptions dispensed per week. Labor inputs are hours worked by each type of pharmacist. Applying our new estimator, we can reject the null hypothesis that independent pharmacies and chains have the same technology. Estimates for independent pharmacies are somewhat noisy but do not suggest that there is a large difference between managers and regular employees. Estimates for chains suggest that managers are more productive than regular employees. We thus conclude that chains seem to improve the effectiveness of managers which may partially explain why they have become the dominant firm type in this industry.

Our second application focuses on education production functions which play a large role in labor and family economics. Here we rely on data from the Child Development Supplement (CDS) of the PSID. We consider two different samples to illustrate the usefulness of our new methods. First, we consider a sample of children who live in married households. Hence, both parental inputs are observed for these children. We find that our matched TSLS estimator produces similar results to the feasible TSLS estimator. We also consider a sample of children from divorced households where father's inputs have to be imputed. Hence, the standard TSLS is no longer feasible, but our matched TSLS can still be applied. We find that our matched TSLS estimator produces insightful estimates in the sample of children from divorced households as well. Both mother's and farther's times are estimated to have positive effects on child quality. However, there are some significant differences between married and divorced parents.

This paper relates to and contributes to the line of literature on production function estimation by proposing a method to handle the problem of partially latent inputs. Our identification strategy is based on strict monotonicity (and the consequent invertibility) in a scalar unobservable, a feature also found in and leveraged by Olley and Pakes (1996) and Levinsohn and Petrin (2003). Their essentially use an auxiliary variable together with an input to control for the unobserved productivity shock: investment with capital in Olley and Pakes (1996) and intermediate inputs with capital in Levinsohn and Petrin (2003). In comparison, we use the *output* with the observed input to pin down the productivity shock. We emphasize that the feature of functional dependence between input variables, which was pointed out by Ackerberg, Caves, and Frazer (2015) as an underlying problem in Olley and Pakes (1996) and Levinsohn and Petrin (2003), in fact forms the basis of our imputation strategy. Beyond these conceptual linkages, our paper has very different focuses from these papers cited above: they focus more on the dynamic nature of capital inputs, and do not use input prices for instruments as discussed in Griliches and Mairesse (1998).

Lastly, we should point out that this paper is both conceptually and technically different from previous work on missing data in linear regression and, more generally, GMM estimation settings, such as Rubin (1976), Little (1992), Robins, Rotnitzky, and Zhao (1994), Wooldridge (2007), Graham (2011), Chaudhuri and Guilkey (2016), Abrevaya and Donald (2017) and McDonough and Millimet (2017). This line of literature usually exploits two types of conditions: first, observations with no missing data occur with positive probability, and second, data are "missing at random" (potentially with conditioning). Neither condition is satisfied in our setting: every observation contains missing data, and missing can be correlated with other observables as well as the unobserved productivity shock. Instead, we rely on monotonicity (and invertibility) in our production function model to identify and impute the latent input.

The rest of the paper is organized as follows. Section 2 discusses the problems associated with estimating the production functions with partially latent input variables. It introduces our new matched TSLS estimator and derives its asymptotic properties. Section 3 reports the results from a variety of different Monte Carlo experiments. Section 4 introduces our first application focusing on the production functions of pharmacies. It discusses our data sources and presents our main empirical findings. Section 5 discusses our second application which deals with education production functions. Finally, Section 6 presents our conclusions.

2 Identification and Estimation

2.1 Model and Identification

Our starting point is a model of N firms, each consisting of one *manager* and one *employee*. For each firm i, let H_{1i} denote the number of hours worked by the manager, H_{2i} denote the hours worked by the employee, and Q_i denote the output. Throughout the paper we use lower-case letters to denote denote the logarithm of the corresponding variables in capital letters, e.g. $q_i = \log Q_i$.

Each firm i operates in one of M competitive local labor markets, indexed by m. Let $m(i)$ denote the market in which firm i is active. The firm must pay the standing local market wages for managers and employees in market m, denoted by W_{1m} and W_{2m} respectively. Wages differ across markets for exogenous reasons such as differences in the costs of living. Heterogeneity of wages across local labor markets will be important for our identification and estimation strategy as discussed in detail below. Often times we will abuse notation and write W_{1i} and W_{2i} for the wages paid by a specific firm i, with the understanding that $W_{1i} = W_{1m(i)}$ and $W_{2i} = W_{2m(i)}$ for any firm i located in market m.

We assume that all firms face a common output price in all markets, i.e., drugs all cost the same. Effectively, we ignore differences in output prices across local labor markets, which is a reasonable assumption if transportation costs are low. For simplicity, we normalize the output price to be one.[7]

We are interested in estimating the production function that relates the output Q_i to the hours worked H_{1i} and H_{2i}, but we are faced with the following data structure with partially latent inputs:

Assumption 1 (Observability)**.**

(i) Either (Q_i, H_{1i}) or (Q_i, H_{2i}), but never both, is observed for each firm i.

(ii) (W_{1m}, W_{2m}) are observed in each local market m.

These structures often arise when the data is collected at the individual or worker-level and only contains a limited amount of firm-level information. These types of unstructured data sets are becoming increasingly more prevalent in empirical work. Below we provide two applications that share this data structure. Our first application focuses on the retail market for pharmaceutical drugs. The data is based on in the National Pharmacist Workforce Survey. This is a survey that is conducted at the employee level. Hence, only one person, either a manager or an employee, is surveyed for each pharmacy. The surveyed person provides store-level information (e.g. output) and her own individual-level information (e.g. weekly worked hours). Our second application is focuses on education production functions and views the household as the relevant unit of analysis. When children have divorced parents it is quire natural that data on the divorced spouse is often not available. Hence, a similar data structure arises in that case. As we discuss in the conclusions of this paper, there are a number of other potential applications where similar data structures arise.

Consider the following production function with Cobb-Douglas technology:

$$q_i = \alpha_0 + \alpha_1 h_{1i} + \alpha_2 h_{2i} + u_i + \epsilon_i \tag{1}$$

where $\alpha := (\alpha_0, \alpha_1, \alpha_2)$ is a vector of unknown parameters of interest with $\alpha_1, \alpha_2 \geq 0$. As in Olley and Pakes (1996), both u_i and ϵ_i are unobserved in the data; however, u_i is

[7] Note that it is straightforward to extend our model to allow different output prices in local markets, as we discuss below.

a "productivity" shock observed (or predicted) by firm i at the time of its input choice decisions (over h_{1i} and h_{2i}), while ϵ_i is a measurement error term that does not enter into firm i's decision problem. We will discuss more general functional forms in Section 2.3.

Let $h_1(u_i, w_i)$ and $h_2(u_i, w_i)$ denote firm i's input choices under local wages $w_i := (w_{1i}, w_{2i})$ and productivity shock u_i. We now impose the following assumption on the monotonicity of the input choice rules h_1 and h_2 with respect to the unobserved productivity shock:

Assumption 2 (Strict Monotonicity)**.** *$h_1(u, w)$ and $h_2(u, w)$ are both strictly increasing in u, for any values of $w = (w_1, w_2)$.*

Assumption 2 is a standard assumption that underlies most, if not all, existing approaches of production function estimation in one way or another. See, for example, Griliches and Mairesse (1998) and Ackerberg, Caves, and Frazer (2015) for reviews of the relevant literature. Note that Assumption 2 is only "sensible" if $\alpha_1, \alpha_2 > 0$ (even though we do not need a formal assumption that $\alpha_1, \alpha_2 > 0$).

Assumption 3 (Wages as IV)**.** *Write $w_i := (w_{1i}, w_{2i})'$ and $h_i := (h_{1i}, h_{2i})'$.*

(i) Relevance: $\mathbb{E}\left[w_i h_i'\right]$ has full rank.

(ii) Exogeneity: $\mathbb{E}\left[u_i | w_i\right] = 0$.

Here we are following a strategy discussed in Griliches and Mairesse (1998) and assume that local wages are valid instruments.[8] Assumptions 2 and 3(i) are automatically satisfied if firms *optimally* choose inputs according to the production function (1), in which case h_1 and h_2 are characterized by the relevant *first-order conditions* and have simple closed-form formulas that are linear and increasing in u (and in w_1, w_2). We note that the problem of partially latent inputs is less relevant in that case, since the "reduced-form" regression of the observed inputs on the exogenous wages w_i will indirectly recover the production function parameters α. This corresponds to the "duality approach" to production function estimation as discussed in detail in Griliches and Mairesse (1998). An attractive feature of our approach is also that we can test whether inputs are optimally chosen. If we reject the null hypothesis that inputs are optimal, our estimator is still feasible while duality estimators are not.[9]

[8] In addition to the exogeneity condition in Assumption 3(ii), we implicitly require through Assumption 3(i) that there is sufficient variation in local wages (e.g., the number of markets M must be at least 3.)

[9] See Appendix A for details on how to implement this test.

Lastly, we require the following exogeneity condition on the measurement error term ϵ_i, in particular with respect to whether h_{1i} or h_{2i} is observed. To state the assumption, define

$$d_i := \begin{cases} 1, & \text{if } H_{1i} \text{ is observed and } H_{2i} \text{ is latent,} \\ 2, & \text{if } H_{2i} \text{ is observed and } H_{1i} \text{ is latent.} \end{cases}$$

Assumption 4 (Measurement Error)**.** $\mathbb{E}\left[\epsilon_i \middle| w_i, h_i, d_i\right] = 0$.

We note that $\mathbb{E}\left[\epsilon_i \middle| w_i, h_i\right] = 0$ is a standard assumption in the literature without the latent input problem. Furthermore, it is worth noting that Assumption 4 is much weaker than the standard "missing-at-random (MAR)" assumption imposed in the literature that focuses on missing data. Here we are simply requiring that ϵ_i is a "measurement error" term that is independent of the covariates (including the indicator for missing variables), but do not impose any restriction on the dependence structure between d_i and (w_i, h_i).

We now present our main identification result, followed by a detailed explanation of our identification strategy.

Theorem 1. *Under model 1 and Assumptions 1-4, both the latent inputs (realizations of* h_{2i} *or* h_{1i}*) and the production function parameters* α *are identified.*

The starting point of our approach is the reduced form of our model with the measurement error term:

$$q_i = \overline{q}\left(u_i, w_i\right) + \epsilon_i \tag{2}$$

where

$$\overline{q}\left(u_i, w_i\right) := \alpha_0 + \alpha_1 h_1\left(u_i, w_i\right) + \alpha_2 h_2\left(u_i, w_i\right) + u_i,$$

which is strictly increasing in u_i given Assumption 2.

First, consider two firms i and j within the same market m so that $w_{1i} = w_{1j}$ and $w_{2i} = w_{2j}$. Suppose that h_{1i}, h_{1j} are observed for i and j while h_{2i}, h_{2j} are unobserved. If these firms have the same value of manager inputs $h_{1i} = h_{1j}$, then it must also be true that they have the same value of the productivity shock, i.e.,

$$u_i = h_1^{-1}\left(h_{1i};\ w_i\right) = h_1^{-1}\left(h_{1j};\ w_i\right) = u_j,$$

where $h_1^{-1}\left(\cdot\,; w_i\right)$ is the inverse of $h_1\left(\cdot, w_i\right)$. This further implies that

$$\overline{q}\left(u_i, w_i\right) = \overline{q}\left(u_j, w_j\right),$$

and hence if we take an average of q_i and q_j,

$$\frac{1}{2}\left(q_i + q_j\right) = \overline{q}\left(w_i, u_i\right) + \frac{1}{2}\left(\epsilon_i + \epsilon_j\right). \tag{3}$$

we are essentially averaging out the variations in ϵ.[10]

Formally, define $\gamma_1(c)$ as the expected output of firm i conditional on the event that h_{1i} is observed ($d_i = 1$) to have some given value c, i.e.,

$$\gamma_1\left(c; w_i\right) := \mathbb{E}\left[\left. q_i \right| w_i,\ d_i = 1,\ h_{1i} = c\right]. \tag{4}$$

Clearly, γ_1 is directly identified from data given Assumption 1, and can be nonparametrically estimated later on. Taking a closer look at γ_1, we have, by equation (2), Assumption 2 and Assumption 4,

$$\begin{aligned}\gamma_1\left(c; w\right) &= \mathbb{E}\left[\left.\overline{q}\left(w_i, u_i\right) + \epsilon_i\right| w_i = w,\ d_i = 1,\ h_1\left(w_i, u_i\right) = c\right] = \overline{q}\left(h_1^{-1}\left(c; w\right), w\right) \\ &= \alpha_0 + \alpha_1 c + \alpha_2 h_2\left(h_1^{-1}\left(c; w\right), w_i\right) + h_1^{-1}\left(c; w\right),\end{aligned} \tag{5}$$

which is a direct formalization of the intuition in equation (3). By conditioning on w_i and a particular *observed value* of $h_{1i} = c$, we are effectively conditioning on the *unobserved* productivity shock u_i, and aggregation across firms allows us to average out the measurement errors and obtain a quantity that is implicitly a function of the productivity shock $u_i = h_1^{-1}\left(c;\ w_i\right)$.

We observe that $\gamma_1(c; w)$ will also be strictly increasing in c, since

$$\frac{\partial}{\partial c}\gamma_1\left(c; w\right) = \alpha_1 + \alpha_2 \frac{\partial}{\partial u} h_2\left(h_1^{-1}\left(c\right), w\right) \frac{1}{\frac{\partial}{\partial u} h_1\left(h_1^{-1}\left(c\right), w\right)} + \frac{1}{\frac{\partial}{\partial u} h_1\left(h_1^{-1}\left(c\right), w\right)} > 0 \tag{6}$$

as $\alpha_1, \alpha_2 \geq 0$ and $\frac{\partial}{\partial u}h_1, \frac{\partial}{\partial u}h_2 > 0$ by Assumption 2. Similarly, we can define $\gamma_2(c; w) := \mathbb{E}\left[\left. q_i \right| w_i = w,\ d_i = 2,\ h_{2i} = c\right]$, which is also strictly increasing in c.

The basic idea behind our identification strategy is then to conditionally "match" observations on the event that

$$\gamma_1\left(c_1; w\right) = \gamma_2\left(c_2; w\right) \tag{7}$$

for some c_1, c_2 at given levels of local wages w.

For example, if again we first consider firms within the same market (so that w is constant), then equation (7) involves two separate (conditionally) expected output levels, one (γ_1) for firms whose h_{1i} is observed , and the other (γ_2) for firms whose h_{2i} is observed. When these two expected output levels are equalized as in equation (7), we can infer that the underlying

[10] As a matter of fact, we can directly match on output q if there is no measurement error in output.

productivity shock u must be the same across all firms with either $h_{1i} = c_1$ observed or $h_{2i} = c_2$ observed, since by equations (3) and (5) we know

$$h_1^{-1}(c_1;\ w) = h_2^{-1}(c_2;\ w) =: \overline{u}$$

which also pins down the latent inputs via:

$$\begin{aligned} h_{2i} &= h_2(w, \overline{u}), \quad \text{for } d_i = 1, \\ h_{1i} &= h_1(w, \overline{u}), \quad \text{for } d_i = 2. \end{aligned}$$

Formally, the latent inputs can be identified as

$$\begin{aligned} h_{2i} &= \gamma_2^{-1}(\gamma_1(h_{1i}; w_i); w_i), \quad \text{for } d_i = 1, \\ h_{1i} &= \gamma_1^{-1}(\gamma_2(h_{2i}; w_i); w_i), \quad \text{for } d_i = 2, \end{aligned} \tag{8}$$

since h_{1i}, h_{2i} are observed for $d_i = 1, 2$ respectively, and γ_1, γ_2 are nonparametrically identified functions.

With the latent inputs identified, we are back to the production function equation (1) without latent inputs, where α is identified under the standard IV condition in Assumption 3.

Alternatively we can consider the following equation which uses the conditional expected output as the dependent variable:

$$\overline{q}_i = \alpha_0 + \alpha_1 h_{1i} + \alpha_2 h_{2i} + u_i, \tag{9}$$

where $\overline{q}_i := \overline{q}(u_i, w_i) = \gamma_1(h_{1i}, w_i) = \gamma_2(h_{2i}, w_i)$. Again, α is identified under Assumption 3.

2.2 Estimation

Our proof of identification is constructive. Hence, it can be used to derive a sequential estimation procedure of the parameters of interest. Specifically, our estimator can be obtained through the following three steps.

Step 1 (Nonparametric Regression): obtain an estimator $\hat{\gamma}_1$ of γ_1 by nonparametrically regressing q_i on h_{1i} and w_i, among firms whose h_{1i} is observed (i.e., $d_i = 1$). Similarly obtain an estimator $\hat{\gamma}_2$ of γ_2.

Step 2 (Imputation): impute latent inputs by plugging the nonparametric estimators $\hat{\gamma}_1, \hat{\gamma}_2$

into equation (8), i.e.,

$$\hat{h}_{2i} = \hat{\gamma}_2^{-1}\left(\hat{\gamma}_1\left(h_{1i}; w_i\right); w_i\right), \quad \text{for } d_i = 1,$$
$$\hat{h}_{1i} = \hat{\gamma}_1^{-1}\left(\hat{\gamma}_2\left(h_{2i}; w_i\right); w_i\right), \quad \text{for } d_i = 2.$$

Step 3 (IV Regression): run either of the following two IV regressions:

- (i) Estimate equation (1) with w_i as IVs, i.e.,

  $$\hat{\alpha} := \left(\frac{1}{n}\sum_{i=1}^{n}\overline{w}_i\tilde{h}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\overline{w}_i q_i\right)$$

  where $\overline{w}_i := \left(1, w_{1i}, w_{2i}\right)'$ and

  $$\tilde{h}_i := \begin{cases} \left(1, h_{1i}, \hat{h}_{2i}\right)', & \text{for } d_i = 1, \\ \left(1, \hat{h}_{1i}, h_{2i}\right)', & \text{for } d_i = 2. \end{cases}$$

- (ii) Estimate (9) with w_i as IVs and the expected output $\overline{q}_i$ replaced by its plug-in estimator

  $$\tilde{q}_i := \begin{cases} \hat{\gamma}_1\left(h_{1i}, w_i\right), & \text{for } d_i = 1, \\ \hat{\gamma}_2\left(h_{2i}, w_i\right), & \text{for } d_i = 2, \end{cases}$$

  i.e.,

  $$\hat{\alpha}^* := \left(\frac{1}{n}\sum_{i=1}^{n}\overline{w}_i\tilde{h}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\overline{w}_i\tilde{q}_i\right).$$

We now derive the consistency and the asymptotic normality of $\hat{\alpha}$ and $\hat{\alpha}^*$ under the following additional regularity assumptions.

Assumption 5 (Finite Error Variances)**.** $\mathbb{E}\left[u_i^2\middle| w_i\right] < \infty$ *and* $\mathbb{E}\left[\epsilon_i^2\middle| w_i, h_i, d_i\right] < \infty$*.*

Assumption 6 (Strong Monotonicity)**.** *The first derivative of* $\gamma_k(\cdot, w)$ *is uniformly bounded away from zero, i.e., for any* c, w*,*

$$\frac{\partial}{\partial c}\gamma_k\left(c; w\right) > \underline{c} > 0.$$

In view of equation (6), Assumption 6 is satisfied if either $\alpha_1, \alpha_2 > 0$ or $\frac{\partial}{\partial u}h_1, \frac{\partial}{\partial u}h_1$ are uniformly bounded *above* by a finite constant. Assumption 6 is needed to ensure that $\hat{\gamma}_k^{-1}(\cdot, w)$ is a good estimator of $\gamma_k^{-1}(\cdot, w)$ provided that the first-stage nonparametric estimator $\hat{\gamma}_k$ is consistent for γ_k, i.e. the two inverse functions need to be well-behaved.

Assumption 7 (First-Stage Estimation)**.**

(i) Donsker property: $\gamma_1, \gamma_2 \in \Gamma$, which is a Donsker class of functions with uniformly bounded first and second derivatives, and $\hat{\gamma}_1, \hat{\gamma}_2 \in \Gamma$ with probability approaching 1.

(ii) First-stage convergence: $\|\hat{\gamma}_k - \gamma_k\| = o_p\left(N^{-\frac{1}{4}}\right)$ for $k = 1, 2$.

Assumption 7(i) is guaranteed if γ_1, γ_2 satisfy certain smoothness condition, e.g. γ_k possesses uniformly bounded derivatives up to a sufficiently high order. 7(ii) is required so that the final estimator of the production function parameters α can converge at the standard parametric ($\sqrt{N}$) rate despite the slower first-step nonparametric estimation of γ_1, γ_2.

Finally, we state another technical assumption that captures how the first-stage nonparametric estimation of γ_1, γ_2 influences the final semiparametric estimators $\hat{\alpha}$ or $\hat{\alpha}^*$ through the functional derivatives of the residual functions with respect to γ_1, γ_2. Assumption 8 below, based on Newey (1994), provides an explicit formula for the asymptotic variance of $\hat{\alpha}$ and $\hat{\alpha}^*$ that does not depend on the particular forms of first-stage nonparametric estimators.

To state Assumption 8, write $z_i := (q_i, w_i, h_i, d_i)$, $\gamma := (\gamma_1, \gamma_2)$, the residual functions

$$g\left(z_i, \tilde{\alpha}, \tilde{\gamma}\right) := \begin{cases} \overline{w}_i \left(q_i - \tilde{\alpha}_0 - \tilde{\alpha}_1 h_{1i} - \tilde{\alpha}_2 \tilde{\gamma}_2^{-1}\left(\tilde{\gamma}_1\left(h_{1i}\right)\right)\right) & \text{for } d_i = 1, \\ \overline{w}_i \left(q_i - \tilde{\alpha}_0 - \tilde{\alpha}_2 h_{2i} - \tilde{\alpha}_1 \tilde{\gamma}_1^{-1}\left(\tilde{\gamma}_2\left(h_{2i}\right)\right)\right) & \text{for } d_i = 2. \end{cases}$$

$$g^*\left(z_i, \tilde{\alpha}, \tilde{\gamma}\right) := \begin{cases} \overline{w}_i \left(\tilde{\gamma}_1\left(h_{1i}\right) - \tilde{\alpha}_0 - \tilde{\alpha}_1 h_{1i} - \tilde{\alpha}_2 \tilde{\gamma}_2^{-1}\left(\tilde{\gamma}_1\left(h_{1i}\right)\right)\right) & \text{for } d_i = 1, \\ \overline{w}_i \left(\tilde{\gamma}_2\left(h_{2i}\right) - \tilde{\alpha}_0 - \tilde{\alpha}_2 h_{2i} - \tilde{\alpha}_1 \tilde{\gamma}_1^{-1}\left(\tilde{\gamma}_2\left(h_{2i}\right)\right)\right) & \text{for } d_i = 2 \end{cases}$$

for generic $\tilde{\alpha}, \tilde{\gamma}$, and

$$g\left(z_i, \tilde{\gamma}\right) := g\left(z_i, \alpha, \tilde{\gamma}\right), \quad g^*\left(z_i, \tilde{\gamma}\right) := g^*\left(z_i, \alpha, \tilde{\gamma}\right),$$

at the true α. The two versions g, g^* of residual functions correspond to the two versions of estimators $\hat{\alpha}, \hat{\alpha}^*$ in Step 3 (i) and (ii): g corresponds to $\hat{\alpha}$, where the raw observed outputs q_i are used in final IV regression, while g^* corresponds to $\hat{\alpha}^*$, where estimates $\tilde{q}_i$ of the conditionally expected output $\overline{q}_i = \gamma_1\left(h_{1i}\right) = \gamma_2\left(h_{2i}\right)$ are used instead. Define the pathwise functional derivative of g at γ along direction τ by

$$G\left(z_i, \tau\right) := \lim_{t \to 0} \frac{1}{t}\left[g\left(z_i, \gamma + t\tau\right) - g\left(z_i, \gamma\right)\right]$$

and similarly define $G^*\left(z_i, \tau\right)$ for g^*. Then, as in Newey (1994), analytical calculation of G and G^* leads to the following characterization of the influence terms from the first-stage estimation[11]:

$$\varphi(z_i) := -\left(\lambda_1 \frac{\alpha_2}{\gamma_2'} - \lambda_2 \frac{\alpha_1}{\gamma_1'}\right)(\mathbb{1}\{d_i = 1\} - \mathbb{1}\{d_i = 2\}).$$

$$\varphi^*(z_i) := \left[\lambda_1\left(1 - \frac{\alpha_2}{\gamma_2'}\right) + \lambda_2 \frac{\alpha_1}{\gamma_1'}\right]\mathbb{1}\{d_i = 1\} + \left[\lambda_1 \frac{\alpha_2}{\gamma_2'} + \lambda_2\left(1 - \frac{\alpha_1}{\gamma_1'}\right)\right]\mathbb{1}\{d_i = 2\}$$

where γ_k' denotes $\frac{\partial}{\partial h_k}\gamma_k(h_{ki}; w_i)$, λ_1 stands for

$$\lambda_1(h_i, w_i) := \mathbb{E}[\mathbb{1}\{d_i = 1\} | h_i, w_i]$$

i.e., the conditional probability of observing h_{1i}, and $\lambda_2 := 1 - \lambda_1$. We are now ready to present the following assumption, which essentially states that the expected error induced by the first-stage estimation is asymptotically equivalent to sample averages of $\varphi(z_i)\overline{w}_i\epsilon_i$ and $\varphi^*(z_i)\overline{w}_i\epsilon_i$.

Assumption 8 (Asymptotic linearity). *(i) Suppose*

$$\int G(z, \hat{\gamma} - \gamma)\, d\mathbb{P}(z) = \frac{1}{N}\sum_{i=1}^{N}\varphi(z_i)\overline{w}_i\epsilon_i + o_p\left(N^{-\frac{1}{2}}\right).$$

(ii) Suppose

$$\int G^*(z, \hat{\gamma} - \gamma)\, d\mathbb{P}(z) = \frac{1}{N}\sum_{i=1}^{N}\varphi^*(z_i)\overline{w}_i\epsilon_i + o_p\left(N^{-\frac{1}{2}}\right).$$

We emphasize that Assumptions 7 and 8 are standard assumptions widely imposed in the semiparametric estimation literature, which can be satisfied by many kernel or sieve first-stage estimators under a variety of conditions. See Newey (1994), Newey and McFadden (1994) and Chen, Linton, and Van Keilegom (2003) for references. In Assumption 9 below, we also provide an example of lower-level conditions that replace Assumptions 7 and 8 when we use the Nadaraya-Watson kernel estimator in the first-stage nonparametric regression.

The next theorem establishes the asymptotic normality of $\hat{\alpha}$ and $\hat{\alpha}^*$.

Theorem 2 (Asymptotic Normality). *Suppose Assumptions 1-7 hold and let* $\Sigma_{wh} := \mathbb{E}\left[\overline{w}_i\overline{h}_i'\right]$.

(i) With Assumption 8(i),

$$\sqrt{N}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma),$$

where $\Sigma := \Sigma_{wh}^{-1}\Omega\Sigma_{hw}^{-1}$ *and*

$$\Omega := \mathbb{E}\left[\overline{w}_i\overline{w}_i'\left(u_i^2 + [1 + \varphi(z_i)]^2\epsilon_i^2\right)\right].$$

[11] See the proof of Theorem 2 for details on the calculation.

(ii) With Assumption 8(ii),

$$\sqrt{N}\left(\hat{\alpha}^{*}-\alpha^{*}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\Sigma^{*}\right),$$

where $\Sigma^{*}:=\Sigma_{wh}^{-1}\Omega^{*}\Sigma_{hw}^{-1}$ *and*

$$\Omega^{*}:=\mathbb{E}\left[\overline{w}_{i}\overline{w}_{i}^{\prime}\left(u_{i}^{2}+\varphi^{*}\left(z_{i}\right)^{2}\epsilon_{i}^{2}\right)\right].$$

(iii) $\Omega-\Omega^{*}$ *is positive definite, i.e.,* $\hat{\alpha}^{*}$ *is asymptotically more efficient than* $\hat{\alpha}$.

We note that, if the latent inputs were observed and the first-step nonparametric regression were not required, the asymptotic variance of standard IV estimator of α would be given by $\Sigma_{wh}^{-1}\mathrm{Var}\left(\overline{w}_{i}\left(u_{i}+\epsilon_{i}\right)\right)\Sigma_{hw}^{-1}$. Hence, the presence of the additional term $\delta\left(z_{i}\right)$ in Ω captures the effect of the first-step nonparametric regression on the asymptotic variance of $\hat{\alpha}$.

Theorem 2(iii) is intuitive: the error term for the IV regression with the raw output q_i as the left-hand-side variable is $u_i+\epsilon_i$, which has a larger variance than the corresponding error term u_i, if the conditionally expected output $\overline{q}_i$ is used instead. Even though we do not observe $\overline{q}_i$ and must use an estimator $\tilde{q}_i=\hat{\gamma}_1\left(h_{1i}\right)$ or $\tilde{q}_i=\hat{\gamma}_2\left(h_{2i}\right)$, the impact of the first-stage estimation error (which can be loosely thought as an *average* of ϵ_i across i) is smaller than the impact of ϵ_i itself.

To see this more clearly, first consider the multiplier "$1+\varphi\left(z_i\right)$" in (i): the "1" comes from the one "raw" share of error ϵ_i embedded in each q_i that we use as the outcome variable, while "$\varphi\left(z_i\right)$" essentially captures the share of influence of the first-step estimation error $\hat{\gamma}-\gamma$ due to ϵ_i. Together, we have

$$1+\varphi=\left(1-\lambda_{1}\frac{\alpha_{2}}{\gamma_{2}^{\prime}}+\lambda_{2}\frac{\alpha_{1}}{\gamma_{1}^{\prime}}\right)\mathbb{1}\left\{d_{i}=1\right\}+\left(\lambda_{1}\frac{\alpha_{2}}{\gamma_{2}^{\prime}}+1-\lambda_{2}\frac{\alpha_{1}}{\gamma_{1}^{\prime}}\right)\mathbb{1}\left\{d_{i}=2\right\},$$

while the corresponding multiplier φ^{*} on ϵ_i in (ii) is essentially the same except that "$1-\lambda_{1}\frac{\alpha_{2}}{\gamma_{2}^{\prime}}$" becomes "$\lambda_{1}-\lambda_{1}\frac{\alpha_{2}}{\gamma_{2}^{\prime}}$" and "$1-\lambda_{2}\frac{\alpha_{1}}{\gamma_{1}^{\prime}}$" becomes "$\lambda_{2}-\lambda_{2}\frac{\alpha_{1}}{\gamma_{1}^{\prime}}$". Since $\lambda_1,\lambda_2<1$, the overall multiplier on ϵ_i becomes smaller in magnitude[12]. Essentially, by using the estimated conditional expected output $\tilde{q}_i$, the raw "1" share of ϵ_i in q_i is moved into the first-stage estimation error of $\overline{q}_i$, which is then "averaged" and reduced in magnitude to λ_1 or λ_2, thus leading to smaller overall variance.

Lastly, we emphasize that the efficiency comparison in 1(iii) does not directly relate to the theory of semiparametric efficiency bounds, such as in Ackerberg et al. (2014), which is about asymptotic efficiency of semiparametric estimators under a *given* criterion function. In fact,

[12] Note that $\alpha_1/\gamma_1^{\prime}\leq 1$ and $\alpha_2/\gamma_2^{\prime}\leq 1$ by equation (6).

by Ackerberg et al. (2014), both estimators based on q_i and $\tilde{q}_i$ attain their *corresponding* semiparametric efficiency bounds with respect to their *different criterion functions g and g^**. 1(iii), however, is a comparison *across* the two criterion functions g and g^*, and essentially states that the (asymptotically) efficient estimator under g^* is even more efficient than the efficient estimator under g.

We now present a set of lower-level conditions that replace Assumption 7 when we use the canonical Nadaraya-Watson kernel estimator for the nonparametric regression in Step 1. We emphasize this simply serves as an illustration of Assumptions 7 and 8 and Theorem 2, as our method does not require the use of any specific form of first-step nonparametric estimators. For sieve (series) first-step estimators, similar results can be derived based on, for example, Newey (1994), Chen (2007) and Chen and Liao (2015).

Assumption 9 (Example Conditions for Kernel First Step). *Let $N_k := \sum_{i=1}^{N} \mathbb{1}\{d_i = k\}$ denote the number of firms for which h_{ki} is observed, and let $\hat{\gamma}_k$ be the Nadaraya-Watson kernel estimator of γ_k defined by*

$$\hat{\gamma}_k(x_k) := \frac{\frac{1}{N_k b^3} \sum_{d_i=k} K\left(\frac{x_k - x_{ki}}{b^3}\right) q_i}{\frac{1}{N_k b^3} \sum_{d_i=k} K\left(\frac{x_k - x_{ki}}{b^3}\right)}$$

where $x_{ki} := (h_{ki}, w_{1i}, w_{2i})$ for all i such that $d_i = k$. Suppose the following conditions hold:

(i) $\lambda_1(h_i, w_i) := \mathbb{E}[\mathbb{1}\{d_i = 1\}| h_i, w_i] \in (\epsilon, 1-\epsilon)$ *for all* (h_i, w_i) *for some* $\epsilon > 0$.

(ii) $(h_{1i}, h_{2i}, w_{1i}, w_{2i})$ *has compact support in* $\mathbb{R}^4$ *with joint density* f *that is uniformly bounded both above and below away from zero.*

(iii) $\mathbb{E}[q_i^4] < \infty$ *and* $\mathbb{E}[q_i^4| h_i, w_i] f(h_i, w_i)$ *is bounded.*

(iv) γ_k *is uniformly bounded and has uniformly bounded derivatives up to order* $p \geq 4$.

(v) $K(u)$ *has uniformly bounded derivatives up to order* p, $K(u)$ *is zero outside a bounded set,* $\int K(u)\,du = 1$, $\int u^t K(u)\,du = \mathbf{0}$ *for* $t = 1, ..., p-1$, *and* $\int \|u\|^p |K(u)|\,du < \infty$.

(vi) b *is chosen such that* $\frac{\sqrt{\log N}}{\sqrt{N b^3}} = o\left(N^{-\frac{1}{4}}\right)$ *and* $\sqrt{N} b^p \to 0$.

Assumption 9(i) essentially requires that the proportion of observations with h_{1i} observed and that with h_{2i} observed are both strictly positive, or in other words, the numbers of both types of observations tend to infinity at the same rate of N. This guarantees that we can estimate both γ_1 based on observations with h_{1i} and γ_2 based on observations with

h_{2i} well enough asymptotically. Assumption 9(iv) is the key smoothness condition that will help establish the Donsker property (and a consequent stochastic equicontinuity condition) in Assumption 7(i). Assumption 9(v)(vi) are concerned with the choice of kernel function K and bandwidth parameter b: (v) requires that a "high-order" kernel function (of order p) is used, while (vi) requires that the bandwidth is set (in a so-called "undersmoothed" way) so that the kernel estimator $\hat{\gamma}_k$ converges at a rate faster than $N^{-1/4}$, as required in Assumption 7(ii). The requirement of $p \geq 4$ in (iii) ensures that (vi) is feasible. Together with the additional regularity conditions in (ii)(ii), these conditions ensure that Assumptions 7-8 and are satisfied. See Newey and McFadden (1994, Section 8.3) for additional details.

Proposition 1 (Asymptotic Distributions with Kernel First Step)**.** *Under Assumptions 1-6 and 9 the conclusions of Theorem 2 hold.*

To obtain consistent variance estimators, define

$$\hat{\Omega} := \frac{1}{N}\sum_{i=1}^{N} \overline{w}_i \overline{w}_i' \left[q_i - \tilde{h}_i'\hat{\alpha} + \hat{\varphi}\left(z_i\right)\left(q_i - \tilde{q}_i\right)\right]^2$$

$$\hat{\Omega}^* := \frac{1}{N}\sum_{i=1}^{N} \overline{w}_i \overline{w}_i' \left[\tilde{q}_i - \tilde{h}_i'\hat{\alpha}^* + \hat{\varphi}^*\left(z_i\right)\left(q_i - \tilde{q}_i\right)\right]^2$$

with

$$\hat{\varphi}\left(z_i\right) := -\left(\hat{\lambda}_1 \frac{\hat{\alpha}_2}{\hat{\gamma}_2'} - \hat{\lambda}_2 \frac{\hat{\alpha}_1}{\hat{\gamma}_1'}\right)\left(\mathbb{1}\left\{d_i = 1\right\} - \mathbb{1}\left\{d_i = 2\right\}\right)$$

$$\hat{\varphi}^*\left(z_i\right) := \left[\hat{\lambda}_1\left(1 - \frac{\hat{\alpha}_2}{\hat{\gamma}_2'}\right) + \hat{\lambda}_2 \frac{\hat{\alpha}_1}{\hat{\gamma}_1'}\right]\mathbb{1}\left\{d_i = 1\right\} + \left[\hat{\lambda}_1 \frac{\hat{\alpha}_2}{\hat{\gamma}_2'} + \hat{\lambda}_2\left(1 - \frac{\hat{\alpha}_1}{\hat{\gamma}_1'}\right)\right]\mathbb{1}\left\{d_i = 2\right\}$$

where $\hat{\lambda}_1 := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{d_i = 1\right\}$. Then the asymptotic variance estimators can be obtained as

$$\hat{\Sigma} := S_{w\tilde{h}}^{-1}\hat{\Omega}S_{\tilde{h}w}^{-1}, \quad \hat{\Sigma}^* := S_{w\tilde{h}}^{-1}\hat{\Omega}^*S_{\tilde{h}w}^{-1}$$

with $S_{w\tilde{h}} := \frac{1}{N}\sum_{i=1}^{N} \overline{w}_i \tilde{h}_i'$.

Proposition 2. *In addition to Assumptions 1-6 and 9, suppose furthermore that* $\lambda_1\left(h_i, w_i\right) \equiv \lambda_1 \in (0,1)$. *Then* $\hat{\Omega} \xrightarrow{p} \Omega$ *and* $\hat{\Omega}^* \xrightarrow{p} \Omega^*$.

2.3 Discussion on Generalizations

Additional Instrumental Variables

If additional instruments are available, it is straightforward to incorporate them in the second-stage regression, which will take the form of a two-stage least square estimator instead of an IV regression. Our results will carry over with suitable changes in notation. For example, the asymptotic variance formula for $\hat{\alpha}$ needs to be adapted as

$$\Sigma := \left(\Sigma_{hw}\Sigma_{ww}^{-1}\Sigma_{wh}\right)^{-1}\Sigma_{hw}\Sigma_{ww}^{-1}\Omega\Sigma_{ww}^{-1}\Sigma_{wh}\left(\Sigma_{hw}\Sigma_{ww}^{-1}\Sigma_{wh}\right)^{-1}.$$

Other Parametric Production Functions

Similarly, the log-linearity of the Cobb-Douglas production function only helps us obtain a linear second-stage regression (IV or TSLS), which does not interfere at all with our first-stage nonparametric estimation of γ_1, γ_2 and the imputation of latent inputs h_{1i} or h_{2i} (beyond the monotonicity conditions built in the production function). To see this, consider a potentially nonlinear (parametric) production function of the form

$$q_i = F_\alpha\left(h_{1i}, h_{2i}\right) + u_i + \epsilon_i$$

such that F_α is increasing in h_{1i} and h_{2i}. Provided that Assumptions 2 and 4 still hold, the conditional expectation

$$\gamma_1\left(c, w\right) := \mathbb{E}\left[\left.q_i\right| w_i = w, d_i = 1, h_{1i} = c\right] = F_\alpha\left(c, h_2\left(h_1^{-1}\left(c; w\right); w\right)\right) + h_1^{-1}\left(c; w\right)$$

remains nonparametrically identified and strictly increasing in c, enabling us to carry over our method of latent input imputation without change. The second stage boils down to the estimation of α based on the moment condition $\mathbb{E}\left[w_i\left(q_i - F_\alpha\left(h_{1i}, h_{2i}\right)\right)\right] = \mathbf{0}$, which can be obtained via nonlinear least square (NLLS) regression or more generally GMM. Technically, since GMM estimators are Z-estimators, the corresponding asymptotic theory in Newey and McFadden (1994), on which the proof of Theorem 1 is based, still applies with proper changes in notations.

Nonparametric Production Function

More generally, with any nonparametric production function that is additively separable in u_i and ϵ_i of the form $q_i = F\left(h_{1i}, h_{2i}\right) + u_i + \epsilon_i$ where F is an unknown function that is increasing in h_{1i} and h_{2i}, our imputation method for the latent input still applies *without change*. The only thing that changes is the second-stage nonparametric estimation of F with both h_{1i} and h_{2i} known (or more precisely, with one known and one imputed) based on the moment condition $\mathbb{E}\left[w_i\left(q_i - F\left(h_{1i}, h_{2i}\right)\right)\right] = \mathbf{0}$. The asymptotic theory for this case can be similarly obtained based on theory on nonparametric two-step estimation (e.g. Hahn, Liao, and Ridder, 2018).

3 A Monte Carlo Experiment

In this section, we report the findings of a Monte Carlo study. Table 1 reports the parameter specifications of the Cobb- Douglas production function that we use in our experiments. Here we assume that inputs are optimally chosen by a profit maximizing firm as discussed in detail in Appendix A. These parameters were chosen so that the simulated data are broadly consistent with the descriptive statistics of our first application that we discuss in detail in the next section. For each specification, market size, denoted by L, and number of firms in each market, denoted by I can vary. In particular, we consider the following scenarios: $L =$ 50, 100, 500 and $I =$ 1, 50, 100. For each experiment, we compute the difference between the true parameter value and the sample average of the estimates using 1000 replications (N). This is a measure of the bias our estimator. We also estimate the root mean squared error (RMSE) using the sample standard deviation of our estimates.

Note that our data generating process mechanically implies h_1 and h_2 have a linear relationship with q. We estimate $\gamma_1(c; w_i)$ and $\gamma_2(c; w_i)$ using a second degree polynomial. Not surprisingly, we find that the estimated coefficients on quadratic terms are almost 0. Moreover the interpolated functions γ_1^{-1} and γ_2^{-1} are almost linear.

Table 2 summarizes the performance of two different estimators: TSLS when all inputs are observed as well as the matched TSLS when inputs are partially latent. As we would expect given our asymptotic results, the matched TSLS performs almost as well as the standard TSLS estimator under these ideal sampling conditions. This finding holds for all three different specifications and several choices for the number of firms within a market and the number of local markets.

Next we investigate how our estimator performs when we have a relatively small number of observations in each market. Considering an extreme case, we simulate data for $L = 500$ and $I = 1$. As we only have a single firm in each market, we cannot impute the missing input

Table 1: Monte Carlo Parameter Specification

	Constant Across Specification						Variable Across Specification		
	α_0	α_1	α_2	μ_z	σ_z	$\kappa_{1,2,3,4}$	σ_u	σ_ϵ	σ_η
Spec1	4	0.35	0.25	$\begin{pmatrix} 2.4 \\ 2.1 \end{pmatrix}$	$\begin{pmatrix} 0.05 & 0 \\ 0 & 0.02 \end{pmatrix}$	$\begin{pmatrix} 1.3 \\ 0.3 \\ 0.1 \\ 0.9 \end{pmatrix}$	0.4	0.3	$\begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$
Spec 2	4	0.35	0.25	$\begin{pmatrix} 2.4 \\ 2.1 \end{pmatrix}$	$\begin{pmatrix} 0.05 & 0 \\ 0 & 0.02 \end{pmatrix}$	$\begin{pmatrix} 1.3 \\ 0.3 \\ 0.1 \\ 0.9 \end{pmatrix}$	0.8	0.3	$\begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}$
Spec 3	4	0.35	0.25	$\begin{pmatrix} 2.4 \\ 2.1 \end{pmatrix}$	$\begin{pmatrix} 0.05 & 0 \\ 0 & 0.02 \end{pmatrix}$	$\begin{pmatrix} 1.3 \\ 0.3 \\ 0.1 \\ 0.9 \end{pmatrix}$	0.8	0.3	$\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$

variable using within market information. Instead, we pool observations across markets and estimate conditional expectations conditional on h_1 (or h_2), w_1, and w_2. Table 2 also summarizes the bias and RMSE where $L = 500$ and $I = 1$. We find that the matched TSLS estimator performs almost as well as the standard TSLS estimator that assumes that both inputs are observed.

Finally, we consider the case in which the wage for type j is observed only when we observe the input for type j, i.e. we assume that:

$$(w_{1i}, w_{2i}) = \begin{cases} (w_{1i}^*, missing) & \text{if } H_{1i} \text{ is observed} \\ (missing, w_{2i}^*) & \text{if } H_{2i} \text{ is observed} \end{cases} \tag{10}$$

Since we need to impute missing wages, we assume that true wages are functions of some demand shifters $Z_m \in \mathbb{R}^2$ for the local labor market m and a random error η_i which is assumed to be independent from the demand shifters. Note that this specification allows for correlation between $w_{1m(i)}$ and $w_{2m(i)}$ through Z_m. Specifically, we simulate wages as follows:

$$w_{1i}^* = w_{1m(i)} = \kappa_1 Z_{1m} + \kappa_2 Z_{2m} + \eta_{1i} \tag{11}$$
$$w_{2i}^* = w_{1m(i)} = \kappa_3 Z_{1m} + \kappa_4 Z_{2m} + \eta_{2i}$$

Table 2: Monte Carlo: Different Markets, Observed Wages

	Number of	Number of		TSLS		Matched TSLS	
Param	Markets	Firms	Spec	Bias	RMSE	Bias	RMSE
α_0	50	50	1	0.001	0.001	0.000	0.001
α_0	100	100	1	-0.000	0.000	-0.000	0.000
α_0	50	50	2	0.001	0.002	-0.000	0.002
α_0	100	100	2	-0.000	0.000	0.000	0.001
α_0	50	50	3	0.001	0.002	0.001	0.002
α_0	100	100	3	-0.000	0.000	0.001	0.001
α_0	500	1	1	-0.004	0.003	-0.004	0.003
α_0	500	1	2	-0.014	0.011	-0.015	0.011
α_0	500	1	3	-0.013	0.010	-0.014	0.010
α_1	50	50	1	0.004	0.003	0.003	0.004
α_1	100	100	1	0.000	0.001	0.000	0.001
α_1	50	50	2	0.007	0.010	0.006	0.013
α_1	100	100	2	0.001	0.002	0.001	0.003
α_1	50	50	3	0.006	0.008	0.032	0.015
α_1	100	100	3	0.001	0.002	0.020	0.003
α_1	500	1	1	-0.002	0.015	-0.001	0.016
α_1	500	1	2	-0.000	0.048	0.001	0.052
α_1	500	1	3	-0.007	0.040	-0.006	0.043
α_2	50	50	1	-0.005	0.005	-0.004	0.006
α_2	100	100	1	-0.001	0.001	-0.000	0.001
α_2	50	50	2	-0.010	0.014	-0.010	0.017
α_2	100	100	2	-0.002	0.003	-0.002	0.004
α_2	50	50	3	-0.007	0.011	-0.046	0.021
α_2	100	100	3	-0.001	0.002	-0.029	0.005
α_2	500	1	1	-0.004	0.020	-0.004	0.022
α_2	500	1	2	-0.020	0.068	-0.022	0.073
α_2	500	1	3	-0.009	0.051	-0.010	0.055

To impute the missing wages, we regress the observed wages (w_{1i}, w_{2i}) on the demand shifters (Z_{1m}, Z_{2m}). Using estimated parameters from the regression, we then impute the missing wages.

Table 3: Monte Carlo: Small Markets with Partially Latent Wages

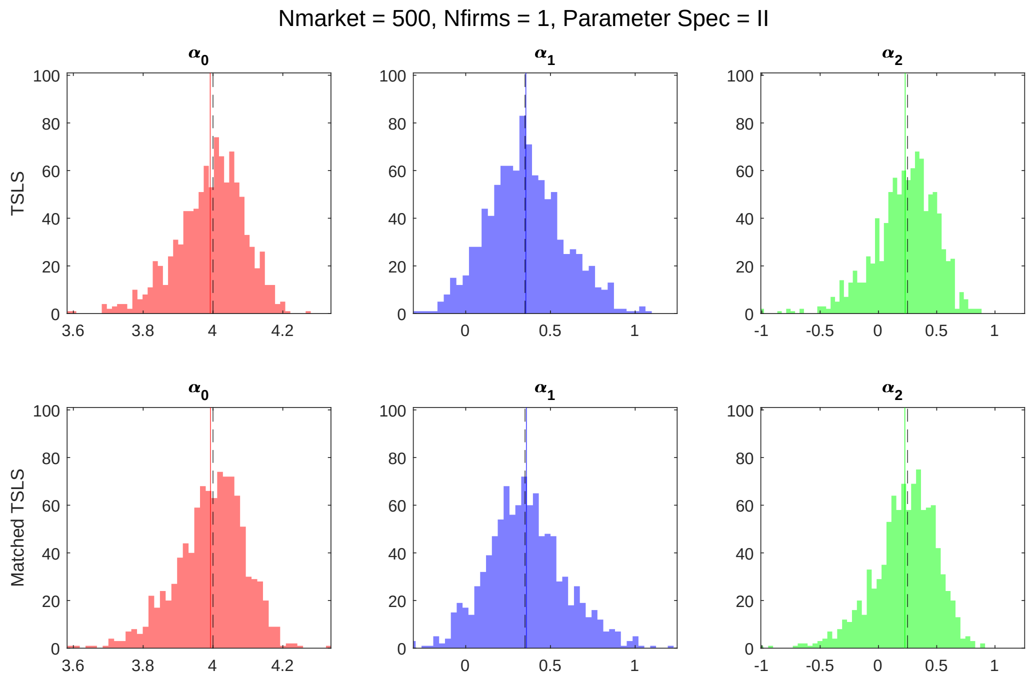
Param	Number of markets	Number of firms	Spec	Standard SLS Bias	Standard SLS RMSE	Matched TSLS Bias	Matched TSLS RMSE
α_0	500	1	1	-0.004	0.003	-0.004	0.003
α_0	500	1	2	-0.008	0.010	-0.007	0.010
α_0	500	1	3	-0.008	0.010	-0.007	0.010
α_1	500	1	1	-0.002	0.015	-0.001	0.016
α_1	500	1	2	0.005	0.054	0.008	0.055
α_1	500	1	3	0.004	0.053	0.008	0.054
α_2	500	1	1	-0.004	0.020	-0.004	0.022
α_2	500	1	2	-0.021	0.072	-0.023	0.075
α_2	500	1	3	-0.020	0.070	-0.023	0.074

Table 3 summarizes the performance of our new estimator together with TSLS estimator. Even if we have a relatively large variance of the imputation errors, such as in Specification 3, our new estimator performs reasonably well.

Figure 1 plots the empirical distribution for the case of specification 2. Overall, we find that the matched TSLS estimator performs almost as well as the standard TSLS estimator.

We conclude that our estimator performs well in all Monte Carlo experiments, even in scenarios that are slightly more general than those considered in Section 2 of the paper. In particular, we do not need to observe both sets of instruments (wages) in the data. We can impute the missing instrument. Next, we evaluate the performance of our estimator in two applications. The first application focuses on pharmacies and studies differences in technology across different types of firms. The second application studies education production functions.

Figure 1: Histograms of Estimated Coefficients With Imputed Wages



4 First Application: Pharmacies

Our first application focuses on the industrial organization of pharmacies. This industry has undergone a dramatic change over the past two decades. An industry that used to be primarily dominated by local independent pharmacies has been transformed by the entry of large chains that operate in multiple markets. An important question is the extent to which this transformation has been driven by technological change that has benefited large chains over smaller independently operated pharmacies. If this is in fact the case, these technological changes may help to explain why this profession has become so popular with females (Goldin and Katz, 2016).

The main data set that we use is the National Pharmacist Workforce Survey of 2000 which is collected by Midwestern Pharmacy Research. The data comes from a cross-sectional survey answered by randomly selected individual pharmacists with active licenses. The data set is composed of two groups of information: information about pharmacists and information about the pharmacy each pharmacist works at.

Information at the pharmacy level includes the type of pharmacy *(Independent* or *Chain)*, the hours of operation per week, the number of pharmacists employed, and the typical number of prescriptions dispensed at the pharmacies per week. The store-level information is provided by an individual pharmacist who works at the pharmacy, thus the quality of the responses may depend on how knowledgeable the person is about the pharmacy. However, considering that most of the pharmacists in our sample are observed to be full-time pharmacists, we think the quality of the firm-level data is fairly high. The number of prescriptions dispensed at the pharmacy is our measure of output. As a consequence, we do not have to use revenue based output measures which could bias our analysis as discussed, for example, in Epple, Gordon, and Sieg (2010).

Table 4: Summary Statistics at the Firm Level: Pharmacies

Firm Type	Number Pharmacists	Emp Size	Operating Hours	Prescriptions per Week	Prescriptions per Hour	Prop Urban	Number of Obs
Indep	$n < 2$	3.15	51.96	778.00	14.94	0.63	50
		(1.41)	(7.08)	(368.95)	(6.54)	(0.39)	
Indep	$2 \leq n < 3$	3.94	56.99	914.40	16.09	0.71	58
		(1.80)	(10.04)	(472.81)	(8.43)	(0.34)	
Indep	$3 \leq n$	4.71	64.24	1252.22	19.44	0.78	36
		(1.44)	(14.15)	(610.61)	(8.75)	(0.32)	
Chain	$n < 2$	1.88	53.50	666.88	12.90	0.81	8
		(0.99)	(8.02)	(278.84)	(6.58)	(0.34)	
Chain	$2 \leq n < 3$	3.25	80.50	1294.68	16.21	0.81	101
		(1.36)	(9.86)	(595.08)	(7.66)	(0.29)	
Chain	$3 \leq n$	5.32	82.82	1765.67	21.43	0.89	79
		(1.63)	(13.67)	(681.57)	(7.87)	(0.20)	

[*] Independent pharmacies: fewer than 10 stores under the same ownership
[*] Chain pharmacies: more than 10 stores under the same ownership
[*] Standard deviations in the parentheses
[*] One part-time pharmacist is counted as 0.5 pharmacist in number of pharmacists
[*] Employment size includes interns and technicians

Table 4 summarizes the means of key variables that are observed at the firm or pharmacy level. After eliminating cases with missing input/output information, we observe 332 pharmacists. Table 4 suggests that there are some pronounced differences between chains and independent pharmacies. Chains are more likely to be located in larger urban areas than independent pharmacies. They also operate longer hours per week. Interestingly, hourly productivity measured by the number of prescriptions per hour is, on average, similar to the independent pharmacies with similar employment size. We explore these issues in more detail below and test whether the different types of pharmacies have access to the same technology.

Table 5: Summary Statistics at the Worker Level: Pharmacists

Firm Type	Position	Number of Pharmacists	Actual Hours	Paid Hours	Hourly Earnings	Number of Obs
Indep	Employee	$n < 2$	40.94	39.28	28.87	9
			(11.61)	(9.60)	(7.64)	
Indep	Employee	$2 \leq n < 3$	33.90	33.03	29.37	29
			(12.01)	(11.14)	(4.09)	
Indep	Employee	$3 \leq n$	31.61	30.95	30.24	28
			(11.62)	(10.96)	(4.93)	
Indep	Manager	$n < 2$	50.02	45.34	30.32	41
			(9.05)	(7.24)	(12.45)	
Indep	Manager	$2 \leq n < 3$	49.45	44.19	28.70	29
			(8.15)	(7.99)	(9.90)	
Indep	Manager	$3 \leq n$	46.50	44.38	30.28	8
			(4.11)	(6.30)	(6.57)	
Chain	Employee	$n < 2$	46.20	43.00	34.70	5
			(2.77)	(4.47)	(2.19)	
Chain	Employee	$2 \leq n < 3$	41.82	39.84	34.13	66
			(5.76)	(4.38)	(3.32)	
Chain	Employee	$3 \leq n$	39.96	37.94	34.03	56
			(8.63)	(7.02)	(3.12)	
Chain	Manager	$n < 2$	45.33	42.00	36.75	3
			(5.03)	(2.65)	(4.43)	
Chain	Manager	$2 \leq n < 3$	44.10	40.50	34.06	35
			(7.02)	(2.58)	(4.90)	
Chain	Manager	$3 \leq n$	43.61	41.43	35.04	23
			(5.41)	(3.41)	(3.59)	

[*] Independent pharmacies: fewer than 10 stores under the same ownership
[*] Chain pharmacies: more than 10 stores under the same ownership
[*] Hourly earnings are computed based on the paid hours, not actual hours
[*] Standard deviations in the parentheses

The data set also collects various information about pharmacists including hours of work, demographics, and household characteristics. Most importantly we observe the position at the pharmacy *(Owner/Manager* or *Employee)*. We treat hours of the manager and hours of

Table 6: Test for Optimality of Inputs

	Independent		Chain	
	H1 Observed	H2 Observed	H1 Observed	H2 Observed
Wald Statistic	5.495	36.914	15.312	26.172
p-value	(0.064)	(0.000)	(0.000)	(0.000)

the employee as the two input factors in our analysis.

Information related to the individual pharmacists is summarized in Table 5. Employee pharmacists at independent pharmacies work fewer hours than the employee pharmacists at chain pharmacies, and hourly earnings are lower than those of the employees at the chains. Pharmacists in managerial positions at independent pharmacies work more hours than do managers at chain pharmacies, but they have lower hourly earnings on average.

We observe only one pharmacy in each local labor market, which is defined as the 5-digit zip code area.[13] Hence, we need to use the version of our estimator that goes across local markets for identification.

We test whether the observed labor inputs are indeed the optimal choice of firms. If the inputs are optimally chosen, the coefficients can be directly estimated from equation 15 in Appendix A. Under the assumption of Cobb-Douglas production, we can test the optimality by jointly testing the null hypothesis of unity of both coefficients on the observed wage and input. Table 6 shows the test result. A formal Wald test rejects the null hypothesis of optimality. Thus the direct inversion of the optimality conditions cannot be applied to estimate the production function; whereas our matched estimator can be applied.

We implement two versions of our matched TSLS estimator using two slightly different matching algorithms. First, we estimate the expectation of output conditional on local wages. Second, we estimate the expectation of output conditional on local demand shifters. We implement these two estimators for the pooled sample and the two subsamples of chains and independent pharmacies. Table 7 summarizes our findings. We report the estimated parameters of the Cobb-Douglas production function as well as the estimated standard errors. In addition, we report standard F-statistics for the first stage of the TSLS estimator to test

[13] We only observe the wage for the observed type. Thus, wages are imputed for the unobserved type using local demand shifters in 5-digit zip code levels and pharmacists' characteristics. We use actual wages for the observed position and imputed wages for both positions together with principal components of local demand shifters as instruments.

for weak instruments. Overall, we find that our instruments are sufficiently strong in most cases.

As discussed in Section 2 of this paper, we can also estimate the production function using expected outputs as the dependent variable. Since the observed output is subject to a measurement error, this semi-parametric estimator offers the potential of some efficiency gains. Table 8 reports our findings for this version of our estimator.

Table 7: Estimation Result

	Independent		Chain	
	Wages	Local Demand Shifters	Wages	Local Demand Shifters
α_0	5.447	4.711	2.504	-1.186
	(0.597)	(0.711)	(1.790)	(4.227)
α_1	0.227	0.230	0.819	1.743
	(0.122)	(0.090)	(0.454)	(1.147)
α_2	0.090	0.285	0.409	0.489
	(0.071)	(0.123)	(0.191)	(0.278)
Nobs	144	144	188	188
First-stage F for h_1	9.320	5.234	11.774	1.747
p-val	(0.000)	(0.000)	(0.000)	(0.074)
First-stage F for h_2	13.648	9.807	3.630	3.528
p-val	(0.000)	(0.000)	(0.000)	(0.000)

Table 8 shows that we estimate all parameters of the production function with good precision. Correcting for potential measurement error by using the expected output as the dependent variable, we achieve similar, maybe even slightly more plausible estimates as shown in Table 8. The estimator that imputes missing values based on local demand conditions performs slightly better in our application than the estimator that just uses wages. This might be because we need to impute wages for missing observations as discussed above. Our results indicate that chains may have a different production function than independent pharmacies. A formal joint hypothesis test rejects the null hypothesis that the coefficients of the production function are the same. The result also suggests that managers may be more effective in chains than independents. A formal Wald test rejects the null hypothesis that the two coefficients that characterize managerial efficiency are the same. Additionally, we do find that chains have a significantly lower residual variance than independents. We thus conclude that chains have different production functions than independent pharmacies

Table 8: Estimation Result With Expected Output As Dependent Variable

	Independent		Chain	
	Wages	Local Demand Shifters	Wages	Local Demand Shifters
α_0	5.857	6.023	3.634	2.992
	(0.331)	(0.369)	(1.060)	(1.598)
α_1	0.163	0.085	0.687	0.815
	(0.057)	(0.042)	(0.268)	(0.432)
α_2	0.047	0.085	0.250	0.297
	(0.051)	(0.065)	(0.105)	(0.106)
Nobs	144	144	188	188
First-stage F for h_1	9.320	5.234	11.774	1.747
p-val	(0.000)	(0.000)	(0.000)	(0.074)
First-stage F for h_2	13.648	9.807	3.630	3.528
p-val	(0.000)	(0.000)	(0.000)	(0.000)

which may partially explain the change in the observed market structure of that industry. However, more research is needed to fully address this question.

5 Second Application: Achievement and Parental Investments

Our second application focuses on the estimation of education production functions. Here we assume that child's achievement q_i is a function of the mother's and father's time inputs, denoted by h_{mi} and h_{fi}. Again, we consider a log-linear Cobb-Douglas specification given by

$$q_i \;=\; \alpha_i \;+\; \alpha_m \, h_{mi} \;+\; \alpha_f \, h_{fi} \;+\; u_i \tag{12}$$

where heterogeneity in intercept is given by:

$$\alpha_i \;=\; x_i' \, \alpha_0 \tag{13}$$

Hence, we assume that baseline productivity α_i varies with family characteristics, such as family income. As before, we can estimate the education production function using TSLS with wages as instruments for inputs as well as our matched TSLS estimator if some inputs are missing.

Our data is based on the four available waves of the Child Development Supplement (CDS). These are the cohorts interviewed in 1997, 2002-3, 2007, and 2014.[14] For these children, we have detailed time usage information of their parents on two days, each of which is randomly selected among weekdays and weekends, respectively. Based on this time diary information, we can construct time inputs for mothers and fathers.[15] The CDS information can be linked to the original PSID survey using the family ID. Hence we have detailed parental information such as education level, household income, and the number of children.

The CDS collects multiple measures of child development including both cognitive and non-cognitive skills. We focus on two important cognitive tests. First, we study the passage comprehension test which assesses reading comprehension and vocabulary among children aged between 6 and 17. Second, we analyze the applied problems test which assesses mathematics reasoning, achievement, and knowledge for children aged between 6 and 17.[16]

We begin by estimating an education production function using the subsample of children who live in married households. Here we observe both mother's and father's inputs directly from the data. We observe 3,236 children with complete inputs and applied problem scores and 2,789 children with complete inputs and reading comprehension scores. Table 9 provides descriptive statistics of the main variables in this sample.

We can estimate the model using the traditional TSLS estimator. We compare these estimates with our matched TSLS which is based on a sample in which we randomly omit one of the two inputs. This exercise allows us to compare the performance of both estimators when there is no latent input problem. Here we restrict our attention to married couples with both spouses living together. We exclude families with more than 5 children. As instruments for time inputs we use education, employment status, hourly wage, age of children. To preserve the representativeness of our sample, we use the child-level survey weight for all analyses. Household labor income is in 10,000 dollars. Table 10 summarizes our findings.

Overall, our empirical findings are reasonable. We find that investments in child quality decrease with the number of children in the family and increase with household income, as expected. Both parental time inputs are positive and typically statistically significant

[14]The CDS 1997 cohort consists of up to 12-year-old children and follows them for 3 waves (1997, 2001, 2007). The CDS 2014 cohort consists of children that were up to 17 years old in 2013.

[15]We exclude families with stepmother and stepfather from our sample.

[16]We also analyzed the letter word test which assesses symbolic learning and reading identification skills. There are also two non-cognitive measures. The externalizing behavioral problem index measures disruptive, aggressive, or destructive behavior. The internalizing behavioral problem index measures expressions of withdrawn, sad, fearful, or anxious feelings.

Table 9: Summary Statistics of CDS Sample

	Married Sample	Divorced Sample
Applied Problem Score (Standardized)	107.58	101.28
	(16.63)	(16.92)
Passage Comprehension Score (Standardized)	105.89	99.48
	(14.77)	(14.49)
Mother's Time Input	20.77	15.18
	(14.32)	(14.06)
Father's Time Input	13.87	4.34
	(11.96)	(13.81)
Total Number of Child In Family	2.17	2.1
	(0.9)	(0.9)
Child's Age At Interview	9.68	11.37
	(4.74)	(4.44)
Total Household Labor Income (in 2011 Dollar)	68941	24158
	(55732)	(28616)
Mother's Age	37.05	37.3
	(7.27)	(6.85)
Father's Age	39.1	38.81
	(7.7)	(8.8)
Mother's Years of Education	13.51	12.92
	(2.57)	(1.97)
Father's Years of Education	13.38	12.97
	(3.21)	(1.9)
Prop of Living With Mother	-	0.88

Table 10: Education Production Function: Married Sample

	Dependent variable:			
	Applied Problems		Passage Comprehension	
	TSLS	matched TSLS	TSLS	matched TSLS
Constant	4.510	4.484	4.321	4.380
	(0.017)	(0.026)	(0.026)	(0.223)
Num Child = 2	−0.011	0.034	−0.051	−0.097
	(0.008)	(0.020)	(0.013)	(0.150)
Num Child = 3+	0.008	0.077	−0.030	−0.059
	(0.009)	(0.026)	(0.014)	(0.152)
Household Labor Inc	0.008	0.006	0.010	0.009
	(0.001)	(0.002)	(0.001)	(0.017)
Mom Hour	0.016	0.027	0.100	0.098
	(0.008)	(0.002)	(0.012)	(0.033)
Dad Hour	0.032	0.021	0.017	0.006
	(0.007)	(0.007)	(0.009)	(0.040)
Nobs	3,236	3,236	2,789	2,789
First-stage F for h_m	61.997	127.295	41.812	58.530
First-stage F for h_f	62.636	117.966	58.654	59.156

Table 11: Education Production Function: Divorced Sample

	Dependent variable:	
	Applied Problems matched TSLS	Passage Comprehension matched TSLS
Constant	4.548	4.529
	(0.127)	(0.061)
Num Child = 2	0.051	0.019
	(0.088)	(0.039)
Num Child = 3+	0.002	−0.015
	(0.112)	(0.066)
Household Labor Inc	−0.013	−0.006
	(0.030)	(0.004)
Mom Hour	0.050	0.037
	(0.044)	(0.015)
Dad Hour	0.010	0.001
	(0.025)	(0.003)
Nobs	785	723
First-stage F for h_m	40.532	35.264
First-stage F for h_f	35.633	33.070

and economically meaningful. Comparing the TSLS with our matched TSLS estimator, we find that the results are remarkably similar, especially for the passage comprehension test. The results for the applied problem test are also encouraging although the differences in the estimates are slightly larger. Qualitatively, we reach the same conclusions with both estimators. We thus conclude that our matched TSLS performs well in this sample.

Next, we consider the subsample that consists of households that self-reported to be either divorced or separated. We exclude "single" households for obvious reasons. In these households, one of the parents is not living in the child's household. We typically do not observe time inputs for these divorced parents. For the applied problem (passage comprehension) score we observe 785 (723) children with the mother's input. There are 103 (92) observations where we have the father's input, which we use for imputation purposes.[17] Note that the standard TSLS is no longer feasible in this subsample because of the latent variable problem. Table 11 summarizes our findings.

[17] Missing instruments for the unobserved spouse are imputed using standard techniques based on the observed spouse's information.

Table 11 shows that the time inputs for fathers and mothers are positive, statistically significant, and economically meaningful. Moreover, the point estimates are similar to the ones we obtained for the married sample reported in Table 10. The main difference is that mother's and father's time inputs are slightly less productive for children from divorced families. In summary, our estimator seems to work well in this application as well and yields plausible and accurate point estimates for most coefficients of interest.

6 Conclusions

We have developed a new method that allows us to estimate production functions when inputs are partially latent. We propose to use a matching algorithm to impute the missing input. The parameters of the production function can then be estimated using a matched TSLS estimator that accounts for the endogeneity of inputs. We have established the asymptotic normality of two versions of our proposed estimators under high-level and low-level conditions, and showed that using the estimated conditionally expected output leads to efficiency gains over using the raw output in the final-step IV regression. We have shown that the estimator performs well in Monte Carlo experiments, even if sampling conditions are not ideal. We consider the case in which we need to go across markets for identification and the situation in which other missing variables need to be imputed as well.

We have also shown that our estimator performs well in two new applications. First, we have estimated the production function of two different types of pharmacies. We consider the case in which pharmacies have two types of labor inputs: managerial and regular pharmacists. Our application is motivated by the observation that pharmacies used to be primarily dominated by local independent firms. More recently, the market structure has been transformed by the entry of large chains that operate in multiple markets. We have studied whether this change in market structure has been driven by differences in technology available to chains and independents. We find some convincing evidence that chains have different technologies than independently operated pharmacies.

Our second application focuses on the estimation of education production functions which play a large role in labor and family economics. We have shown that our matched TSLS estimator produces similar results to the regular TSLS estimator in a sample of children in married households, where both parental inputs are observed. We have also considered a sample of children from divorced households where father's inputs have to be imputed. We find that our matched TSLS estimator produces insightful estimates in that sample as well.

This paper provides ample scope for future research. We discussed in Section 2.3 that our estimator can be extended to allow for functional form assumptions than the standard Cobb-Douglas Case. Moreover, it may be possible to incorporate dynamic input considerations and richer error structure in the production function than we have considered thus far. Finally, we think that our estimator will also turn out to be fruitful in many other applications. In particular, our approach seems to be most promising for firms with a small number of workers who perform different tasks.

References

Abrevaya, J. and S. G. Donald (2017): "A GMM approach for dealing with missing data on regressors," *Review of Economics and Statistics*, 99, 657–662.

Ackerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007): "Econometric tools for analyzing market outcomes," *Handbook of Econometrics*, 6, 4171–4276.

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic efficiency of semiparametric two-step GMM," *Review of Economic Studies*, 81, 919–943.

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

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

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

Chaudhuri, S. and D. K. Guilkey (2016): "GMM with multiple missing variables," *Journal of Applied Econometrics*, 31, 678–706.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

Chen, X. and Z. Liao (2015): "Sieve semiparametric two-step GMM under weak dependence," *Journal of Econometrics*, 189, 163–186.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

Epple, D., B. Gordon, and H. Sieg (2010): "A new approach to estimating the production function for housing," *American Economic Review*, 100, 905–24.

Goldin, C. and L. F. Katz (2016): "A most egalitarian profession: pharmacy and the evolution of a family-friendly occupation," *Journal of Labor Economics*, 34, 705–746.

Graham, B. S. (2011): "Efficiency bounds for missing data models with semiparametric restrictions," *Econometrica*, 79, 437–452.

Griliches, Z. and J. Mairesse (1998): "Production Functions: The Search for Identification," in *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, ed. by S. Strøm, Cambridge University Press, 169–203.

Hahn, J., Z. Liao, and G. Ridder (2018): "Nonparametric two-step sieve M estimation and inference," *Econometric Theory*, 34, 1281–1324.

Hansen, B. E. (2008): "Uniform convergence rates for kernel estimation with dependent data," *Econometric Theory*, 726–748.

Hoch, I. (1955): "Estimation of production function parameters and testing for efficiency," *Econometrica*, 23, 325–26.

——— (1962): "Estimation of production function parameters combining time-series and cross-section data," *Econometrica*, 34–53.

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

Little, R. J. (1992): "Regression with missing X's: a review," *Journal of the American statistical association*, 87, 1227–1237.

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

McDonough, I. K. and D. L. Millimet (2017): "Missing data, imputation, and endogeneity," *Journal of Econometrics*, 199, 141–155.

Mundlak, Y. (1961): "Empirical production function free of management bias," *Journal of Farm Economics*, 43, 44–56.

——— (1963): "Specification and estimation of multiproduct production functions," *Journal of Farm Economics*, 45, 433–443.

Newey, K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica*, 1349–1382.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American statistical Association*, 89, 846–866.

Rubin, D. B. (1973): "Matching to remove bias in observational studies," *Biometrics*, 159–183.

——— (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66, 688.

——— (1976): "Inference and missing data," *Biometrika*, 63, 581–592.

Van Der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

Wooldridge, J. M. (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of econometrics*, 141, 1281–1301.

A The Cobb-Douglas Case with Optimal Inputs

Suppose that firm i chooses inputs optimally by solving the following (expected) profit-maximization problem:

$$\max_{H_{1i},H_{2i}} e^{\alpha_0} H_{1i}^{\alpha_1} \; H_{2i}^{\alpha_2} \; e^{u_i} - W_{1i}H_{1i} - W_{2i}H_{2i}, \tag{14}$$

By the first-order conditions,

$$\begin{aligned}
H_{1i} &= e^{\frac{\alpha_0+u_i}{1-\alpha_1-\alpha_2}} \left(\frac{W_{1i}}{\alpha_1}\right)^{\frac{1-\alpha_2}{\alpha_1+\alpha_2-1}} \left(\frac{W_{2i}}{\alpha_2}\right)^{\frac{\alpha_2}{\alpha_1+\alpha_2-1}} \\
H_{2i} &= e^{\frac{\alpha_0+u_i}{1-\alpha_1-\alpha_2}} \left(\frac{W_{2i}}{\alpha_2}\right)^{\frac{1-\alpha_1}{\alpha_1+\alpha_2-1}} \left(\frac{W_{1i}}{\alpha_1}\right)^{\frac{\alpha_1}{\alpha_1+\alpha_2-1}} \\
Q_i &= e^{\frac{\alpha_0+u_i}{1-\alpha_1-\alpha_2}} \left(\frac{W_{1i}}{\alpha_1}\right)^{\frac{\alpha_1}{\alpha_1+\alpha_2-1}} \left(\frac{W_{2i}}{\alpha_2}\right)^{\frac{\alpha_2}{\alpha_1+\alpha_2-1}} \\
&= e^{\alpha_0+u_i} \left(\frac{\alpha_2 W_{1i}}{\alpha_1 W_{2i}}\right)^{\alpha_2} H_{1i}^{\alpha_1+\alpha_2} = e^{\alpha_0+u_i} \left(\frac{\alpha_1 W_{2i}}{\alpha_2 W_{1i}}\right)^{\alpha_1} H_{2i}^{\alpha_1+\alpha_2}
\end{aligned}$$

In log forms

$$\begin{aligned}
h_1(u_i, w_i) &= \frac{\alpha_0 + (1-\alpha_2)\log\alpha_1 + \alpha_2 \log\alpha_2}{1-\alpha_1-\alpha_2} - \frac{1-\alpha_2}{1-\alpha_1-\alpha_2} w_{1i} - \frac{\alpha_2}{1-\alpha_1-\alpha_2} w_{2i} + \frac{1}{1-\alpha_1-\alpha_2} u_i \\
h_2(u_i, w_i) &= \frac{\alpha_0 + \alpha_1 \log\alpha_1 + (1-\alpha_1)\log\alpha_2}{1-\alpha_1-\alpha_2} - \frac{\alpha_1}{1-\alpha_1-\alpha_2} w_{1i} - \frac{1-\alpha_1}{1-\alpha_1-\alpha_2} w_{2i} + \frac{1}{1-\alpha_1-\alpha_2} u_i \\
\overline{q}(u_i, w_i) &= \frac{\alpha_0 + \alpha_1 \log\alpha_1 + \alpha_2 \log\alpha_2}{1-\alpha_1-\alpha_2} - \frac{\alpha_1}{1-\alpha_1-\alpha_2} w_{1i} - \frac{\alpha_2}{1-\alpha_1-\alpha_2} w_{2i} + \frac{1}{1-\alpha_1-\alpha_2} u_i \\
&= \alpha_0 + \alpha_2 \log(\alpha_2/\alpha_1) + (\alpha_1+\alpha_2)\, h_1(u_i, w_i) + \alpha_2 w_{1i} - \alpha_2 w_{2l} + u_i \\
&= \alpha_0 + \alpha_1 \log(\alpha_1/\alpha_2) + (\alpha_1+\alpha_2)\, h_2(u_i, w_i) - \alpha_1 w_{1i} + \alpha_1 w_{2l} + u_i
\end{aligned}$$

Taking inverses

$$\begin{aligned}
u_i = h_1^{-1}(h_{1i}, w_i) &:= -\left[\alpha_0 + (1-\alpha_2)\log\alpha_1 + \alpha_2\log\alpha_2\right] + (1-\alpha_1-\alpha_2)\, h_{1i} + (1-\alpha_2)\, w_{1i} + \alpha_2 w_{2i} \\
= h_2^{-1}(h_{2i}, w_i) &:= -\left[\alpha_0 + \alpha_1\log\alpha_1 + (1-\alpha_1)\log\alpha_2\right] + (1-\alpha_1-\alpha_2)\, h_{2i} + \alpha_1 w_{1i} + (1-\alpha_1)\, w_{2i}
\end{aligned}$$

Hence,

$$\begin{aligned}
\gamma_1(h_{1i}, w_i) &= \overline{q}\left(h_1^{-1}(h_{1i}, w_i), w_i\right) = -\log\alpha_1 + h_{1i} + w_{1i}, \\
\gamma_2(h_{2i}, w_i) &= \overline{q}\left(h_2^{-1}(h_{2i}, w_i), w_i\right) = -\log\alpha_2 + h_{2i} + w_{2i},
\end{aligned}$$

and

$$\begin{aligned} q_i &= \gamma_1(h_{1i}, w_i) + \epsilon_i = -\log\alpha_1 + h_{1i} + w_{1i} + \epsilon_i \\ &= \gamma_2(h_{2i}, w_i) + \epsilon_i = -\log\alpha_2 + h_{2i} + w_{2i} + \epsilon_i. \end{aligned} \tag{15}$$

It is then evident that α_1 or α_2 can be estimated directly from 15 from the corresponding subsample where h_{1i} or h_{2i} is observed.

B Proofs

B.1 Additional Notation and Lemmas

Notation For each i, we use h_{ji} to denote the *observed* input and use h_{ki} to denote the *latent* input variable for firm i, i.e.

$$h_{ji} = h_{1i},\ h_{ki} = h_{2i},\ \text{for } d_i = 1,$$
$$h_{ji} = h_{2i},\ h_{ki} = h_{1i},\ \text{for } d_i = 2.$$

We write

$$d_{i1} := \mathbb{1}\{d_i = 1\},$$
$$d_{i2} := \mathbb{1}\{d_i = 2\},$$

so that $h_{ji} = d_{i1}h_{1i} + d_{i2}h_{2i}$ while $h_{ki} := d_{i1}h_{2i} + d_{i2}h_{1i}$.

We write $\overline{h}_i := (1, h_{1i}, h_{2i})'$ to denote the true regressor vector. (Recall $\tilde{h}_i$ denotes the same regressor vector with imputed latent input $\hat{h}_{ki}$ in place of h_{ki}.)

Moreover, we suppress the wage variables w in functions such as $\gamma_1(u, w)$ and $\gamma_2(u, w)$, unless it becomes necessary to emphasize the dependence of such functions on w.

Lemma 1. *Under Assumption 6, if* $\|\hat{\gamma}_k - \gamma_k\|_\infty = O_p(a_n)$, *then* $\left\|\hat{\gamma}_k^{-1} - \gamma_k^{-1}\right\|_\infty = O_p(a_n)$ *and* $\left|\hat{h}_{ki} - h_{ki}\right| = O_p(a_n)$.

Proof. By Assumption 6 we have

$$\underline{c}\,|u_1 - u_2| \leq |\gamma_k(u_1) - \gamma_k(u_2)|$$

For any $v \in Range(\gamma_k)$,

$$\left|\hat{\gamma}_k^{-1}(v) - \gamma_k^{-1}(v)\right| \leq \frac{1}{\underline{c}}\left|\gamma_k\left(\hat{\gamma}_k^{-1}(v)\right) - \gamma_k\left(\gamma_k^{-1}(v)\right)\right| = \frac{1}{\underline{c}}\left|\gamma_k\left(\hat{\gamma}_k^{-1}(v)\right) - v\right|$$
$$= \frac{1}{\underline{c}}\left|\gamma_k\left(\hat{\gamma}_k^{-1}(v)\right) - \hat{\gamma}_k\left(\hat{\gamma}_k^{-1}(v)\right)\right| \leq \frac{1}{\underline{c}}\left\|\hat{\gamma}_k - \gamma_k\right\|_\infty = O_p(a_n).$$

Furthermore, observing that

$$\underline{c}\left|\gamma_k^{-1}(v_1) - \gamma_k^{-1}(v_2)\right| \leq \left|\gamma_k\left(\gamma_k^{-1}(v_1)\right) - \gamma_k\left(\gamma_k^{-1}(v_2)\right)\right| = |v_1 - v_2|$$

we have by Assumption 6 and Lemma 1, for $d_i = 1$,

$$\begin{aligned}
\left|\hat{h}_{ki} - h_{ki}\right| &= \left|\hat{\gamma}_j^{-1}(\hat{\gamma}_k(h_{ki})) - \gamma_j^{-1}(\gamma_k(h_{ki}))\right| \\
&= \left|\hat{\gamma}_j^{-1}(\hat{\gamma}_k(h_{ki})) - \gamma_j^{-1}(\hat{\gamma}_k(h_{ki})) + \gamma_j^{-1}(\hat{\gamma}_k(h_{ki})) - \gamma_j^{-1}(\gamma_k(h_{ki}))\right| \\
&\leq \left|\hat{\gamma}_j^{-1}(\hat{\gamma}_k(h_{ki})) - \gamma_j^{-1}(\hat{\gamma}_k(h_{ki}))\right| + \left|\gamma_j^{-1}(\hat{\gamma}_k(h_{ki})) - \gamma_j^{-1}(\gamma_k(h_{ki}))\right| \\
&\leq \left\|\hat{\gamma}_j^{-1} - \gamma_j^{-1}\right\|_\infty + \frac{1}{\underline{c}}\left|\hat{\gamma}_k(h_{ki}) - \gamma_k(h_{ki})\right| \\
&\leq \left\|\hat{\gamma}_j^{-1} - \gamma_j^{-1}\right\|_\infty + \frac{1}{\underline{c}}\left\|\hat{\gamma}_k - \gamma_k\right\|_\infty \\
&= O_p(a_n). \qquad (16)
\end{aligned}$$

□

Lemma 2. *Under Assumption 6:*

(i) The pathwise derivative of γ_k^{-1} w.r.t. γ_k along $\tau_k \in \Gamma$ is given by

$$\nabla_{\gamma_k}\gamma_k^{-1}[\tau_k] := \lim_{t \searrow 0}\frac{(\gamma_k + t\tau_k)^{-1}(v) - \gamma_k^{-1}(v)}{t} = -\frac{\tau_k\left(\gamma_k^{-1}(v)\right)}{\gamma_k'\left(\gamma_k^{-1}(v)\right)}.$$

(ii) The pathwise derivative of $\gamma_k^{-1}(\gamma_j(\cdot))$ w.r.t. γ_j along $\tau_j \in \Gamma$ is given by

$$\begin{aligned}
\nabla_{\gamma_j}\left(\gamma_k^{-1} \circ \gamma_j\right)[\tau_j] &:= \lim_{t \searrow 0}\frac{\gamma_k^{-1}(\gamma_j(x) + t\tau_j(x)) - \gamma_k^{-1}(\gamma_j(x))}{t} \\
&= \left(\gamma_k^{-1}\right)'(\gamma_j(x))\,\tau_j(x) = \frac{1}{\gamma_k'\left(\gamma_k^{-1}(\gamma_j(x))\right)}\tau_j(x).
\end{aligned}$$

(iii) The second-order derivatives have bounded norms:

$$\nabla^2_{\gamma_k}\gamma_k^{-1}[\tau_k][\tau_k] \leq M\left\|\tau_k\right\|^2$$
$$\nabla^2_{\gamma_j}\left(\gamma_k^{-1} \circ \gamma_j\right)[\tau_j][\tau_j] \leq M\left\|\tau_k\right\|^2$$

Proof. (i) and (ii) follow immediately from the definition of pathwise derivatives. See, e.g., Lemma 3.9.20 and 3.9.25 in Van Der Vaart and Wellner (1996) for reference. For (iii),

$$\nabla^2_{\gamma_k}\gamma_k^{-1}[\tau_k][\nu_k] = \frac{\tau_k'\left(\gamma_k^{-1}\right)}{\gamma_k'\left(\gamma_k^{-1}\right)} \cdot \frac{\nu_k\left(\gamma_k^{-1}\right)}{\gamma_k'\left(\gamma_k^{-1}\right)} - \frac{\tau_k\left(\gamma_k^{-1}\right)}{\left[\gamma_k'\left(\gamma_k^{-1}\right)\right]^2}\left[\gamma_k''\left(\gamma_k^{-1}\right) + \frac{1}{\gamma_k'\left(\gamma_k^{-1}\right)}\right]\nu_k\left(\gamma_k^{-1}\right)$$
$$\leq M\|\tau_k\|\|\nu_k\|$$

since $\gamma_k' \geq \underline{c} > 0$ by Assumption 6 and γ'' and τ_k' are uniformly bounded above by Assumption 7(i). Similarly for $\nabla^2_{\gamma_j}\left(\gamma_k^{-1}\circ\gamma_j\right)$. □

Lemma 3. *Writing* $\gamma := (\gamma_1, \gamma_2)$*, the pathwise derivative of* $\gamma_k^{-1}\circ\gamma_j$ *w.r.t.* γ *along* τ *is given by*

$$\nabla_\gamma\left(\gamma_k^{-1}\circ\gamma_j\right)[\tau] := \lim_{t\searrow 0}\frac{(\gamma_k + t\tau_k)^{-1}(\gamma_j(x) + t\tau_j(x)) - \gamma_k^{-1}(\gamma_j(x))}{t}$$
$$= \frac{1}{\gamma_k'\left(\gamma_k^{-1}(\gamma_j(x))\right)}\left[\tau_j(x) - \tau_k\left(\gamma_k^{-1}(\gamma_j(x))\right)\right]$$

Proof. By Lemma 2,

$$\frac{1}{t}\left[(\gamma_k + t\tau_k)^{-1}(\gamma_j(x) + t\tau_j(x)) - \gamma_k^{-1}(\gamma_j(x))\right]$$
$$= \frac{1}{t}\left[(\gamma_k + t\tau_k)^{-1}(\gamma_j(x) + t\tau_j(x)) - \gamma_k^{-1}(\gamma_j(x) + t\tau_j(x))\right]$$
$$+ \frac{1}{t}\left[\gamma_k^{-1}(\gamma_j(x) + t\tau_j(x)) - \gamma_k^{-1}(\gamma_j(x))\right]$$
$$\to \nabla_{\gamma_k}\gamma_k^{-1}[\tau_k](\gamma_j(x)) + \nabla_{\gamma_j}\left(\gamma_k^{-1}\circ\gamma_j\right)[\tau_j]$$
$$= -\frac{\tau_k\left(\gamma_k^{-1}(\gamma_j(x))\right)}{\gamma_k'\left(\gamma_k^{-1}(\gamma_j(x))\right)} + \frac{1}{\gamma_k'\left(\gamma_k^{-1}(\gamma_j(x))\right)}\tau_j(x)$$
$$= \frac{1}{\gamma_k'\left(\gamma_k^{-1}(\gamma_j(x))\right)}\left(\tau_j(x) - \tau_k\left(\gamma_k^{-1}(\gamma_j(x))\right)\right)$$

□

B.2 Proof of Theorem 2(i)

Proof. We verify the conditions in Lemma 5.4 of Newey (1994), or equivalently, Theorems 8.11 of Newey and McFadden (1994).

Recall $z_i := (q_i, w_i, h_i, d_i)$, $\gamma := (\gamma_1, \gamma_2)$ and

$$\begin{aligned} g(z,\hat\alpha,\hat\gamma) =& \overline{w}_i\left(q_i - \hat\alpha_0 - \left(h_{1i}\hat\alpha_1 + \hat\gamma_2^{-1}(\hat\gamma_1(h_{1i}))\hat\alpha_2\right)d_{i1} - \left(h_{2i}\hat\alpha_2 + \hat\gamma_1^{-1}(\hat\gamma_2(h_{2i}))\hat\alpha_2\right)d_{i2}\right) \\ =& \overline{w}_i\left(q_i - \hat\alpha_0 - h_{ji}\hat\alpha_j - \hat\gamma_k^{-1}(\hat\gamma_j(h_{ji}))\hat\alpha_k\right) \\ g(z,\hat\gamma) =& \overline{w}_i\left(q_i - \alpha_0 - \left(h_{1i}\alpha_1 + \hat\gamma_2^{-1}(\hat\gamma_1(h_{1i}))\alpha_2\right)d_{i1} - \left(h_{2i}\alpha_2 + \hat\gamma_1^{-1}(\hat\gamma_2(h_{2i}))\alpha_2\right)d_{i2}\right) \\ =& \overline{w}_i\left(q_i - \alpha_0 - h_{ji}\alpha_j - \hat\gamma_k^{-1}(\hat\gamma_j(h_{ji}))\alpha_k\right) \\ =& \overline{w}_i\left(u_i + \epsilon_i + \left[h_{ki} - \hat\gamma_k^{-1}(\hat\gamma_j(h_{ji}))\right]\alpha_k\right) \end{aligned}$$

Clearly, $\mathbb{E}[g(z_i,\gamma)] = \mathbb{E}[\overline{w}_i(u_i+\epsilon_i)] = 0$ by Assumptions 3 and 4. Moreover, $\frac{1}{N}\sum_{i=1}^N g(z,\hat\alpha,\hat\gamma) = 0$ by the definition of $\hat\alpha$.

Now, define

$$\begin{aligned} G(z_i,\hat\gamma-\gamma) :=& \nabla_\gamma g(z,\gamma)[\hat\gamma-\gamma] \\ =& -\alpha_k\overline{w}_i\nabla_\gamma\left(\gamma_k^{-1}\circ\gamma_j\right)[\hat\gamma-\gamma] \\ =& \frac{-\alpha_k\overline{w}_i}{\gamma_k'\left(\gamma_k^{-1}(\gamma_j(h_{ji}))\right)}\left[(\hat\gamma_j-\gamma_j)(h_{ji}) - (\hat\gamma_k-\gamma_k)\left(\gamma_k^{-1}(\gamma_j(h_{ji}))\right)\right] \\ =& -\frac{\alpha_k\overline{w}_i}{\gamma_k'(h_{ki})}\left[\hat\gamma_j(h_{ji}) - \gamma_j(h_{ji}) - \hat\gamma_k(h_{ki}) + \gamma_k(h_{ki})\right] \text{ since } \gamma_k^{-1}(\gamma_j(h_{ji})) = h_{ki} \\ =& d_{i1}\overline{w}_i\left(-\frac{\alpha_2}{\gamma_2'}\right)(1,-1)\begin{pmatrix}\hat\gamma_1-\gamma_1\\ \hat\gamma_2-\gamma_2\end{pmatrix} + d_{i2}\overline{w}_i\left(-\frac{\alpha_1}{\gamma_1'}\right)(-1,1)\begin{pmatrix}\hat\gamma_1-\gamma_1\\ \hat\gamma_2-\gamma_2\end{pmatrix} \\ =& -\overline{w}_i\left(d_{i1}\frac{\alpha_2}{\gamma_2'} - d_{i2}\frac{\alpha_1}{\gamma_1'}\right)(1,-1)(\hat\gamma-\gamma) \end{aligned} \tag{17}$$

By Lemma 2(iii) and Lemma 3, we deduce

$$\|g(z,\hat\gamma) - g(z,\gamma) - G(z,\hat\gamma-\gamma)\| = O_p\left(\|\hat\gamma-\gamma\|_\infty^2\right) = o_p\left(\frac{1}{\sqrt{N}}\right)$$

given our assumption that $\|\hat\gamma-\gamma\|_\infty = o_p\left(N^{-1/4}\right)$.

Next, the stochastic equicontinuity condition

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N\left(G(z,\hat\gamma-\gamma) - \int G(z,\hat\gamma-\gamma)\,d\mathbb{P}(z)\right) = o_p\left(\frac{1}{\sqrt{N}}\right) \tag{18}$$

is guaranteed by Assumptions 6 and 7. Specifically, $\hat\gamma-\gamma$ belongs to a Donsker class of functions by the smoothness assumption while $1/\gamma_k'(h_{ki}) \le 1/\underline{c}$ guarantees that $G(z_i,\cdot)$ is square-integrable, so that $G(z_i,\cdot)$ is also Donsker and thus (18) holds.

Now, write

$$x_i := (h_i, w_i)$$

so that $z_i = (q_i, x_i, d_i)$. Then we have

$$\begin{aligned}
&\int G\left(z_i, \hat{\gamma} - \gamma\right) \mathbb{P} z_i \\
=& \int -\overline{w}_i \left(d_{i1} \frac{\alpha_2}{\gamma_2^{'}} - d_{i2} \frac{\alpha_1}{\gamma_1^{'}} \right) (1, -1) \left(\hat{\gamma} - \gamma\right) d\mathbb{P}\left(x_i, d_i\right) \\
=& \int -\overline{w}_i \left(\left[\int d_{i1} d\mathbb{P}\left(d_i | x_i\right) \right] \frac{\alpha_2}{\gamma_2^{'}} - \left[\int d_{i2} d\mathbb{P}\left(d_i | x_i\right) \right] \frac{\alpha_1}{\gamma_1^{'}} \right) (1, -1) \left(\hat{\gamma} - \gamma\right) d\mathbb{P} x_i \\
=& \int -\overline{w}_i \left(\lambda_1\left(x_i\right) \frac{\alpha_2}{\gamma_2^{'}} - \lambda_2\left(x_i\right) \frac{\alpha_1}{\gamma_1^{'}} \right) (1, -1) \left(\hat{\gamma} - \gamma\right) d\mathbb{P} x_i
\end{aligned}$$

By Proposition 4 of Newey (1994), with

$$\varphi\left(z_i\right) := -\left(\lambda_1 \frac{\alpha_2 \overline{w}_i}{\gamma_2^{'}} - \lambda_2 \frac{\alpha_1 \overline{w}_i}{\gamma_1^{'}} \right) \left(d_{i1} - d_{i2}\right)$$

we have

$$\overline{w}_i \left(\lambda_1 \frac{\alpha_2}{\gamma_2^{'}} - \lambda_2 \frac{\alpha_1}{\gamma_1^{'}} \right) (1, -1) \begin{pmatrix} d_{i1}\left(q_i - \gamma_1\left(h_{1i}\right)\right) \\ d_{i2}\left(q_i - \gamma_2\left(h_{2i}\right)\right) \end{pmatrix} \equiv \varphi\left(z_i\right) \overline{w}_i \epsilon_i,$$

and by Assumption 8

$$\int G\left(z, \hat{\gamma} - \gamma\right) d\mathbb{P}\left(z\right) = \frac{1}{N} \sum_{i=1}^{N} \varphi\left(z_i\right) \overline{w}_i \epsilon_i + o_p\left(\frac{1}{\sqrt{N}}\right).$$

Hence, Lemma 5.4 of Newey (1994),

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} g\left(z_i, \hat{\gamma}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[g\left(z_i, \gamma\right) + \varphi\left(z_i\right) \overline{w}_i \epsilon_i \right] + o_p\left(1\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Omega\right),$$

where

$$\begin{aligned}
\Omega :=& \mathrm{Var}\left[g\left(z_i, \gamma\right) + \varphi\left(z_i\right) \overline{w}_i \epsilon_i \right] \\
=& \mathbb{E}\left[\overline{w}_i \overline{w}_i^{'} \left(u_i + \left[1 + \varphi\left(z_i\right)\right] \epsilon_i\right)^2 \right] = \mathbb{E}\left[\overline{w}_i \overline{w}_i^{'} \left(u_i^2 + \left[1 + \varphi\left(z_i\right)\right]^2 \epsilon_i^2\right) \right]
\end{aligned}$$

Lastly, by Lemma 1

$$\left| \frac{1}{n} \sum_{i=1}^{n} \overline{w}_i \left(\hat{h}_{1i} - h_{1i}\right) \right| \leq \frac{1}{n} \sum_{i=1}^{n} \left|\overline{w}_i\right| \left|\hat{h}_{1i} - h_{1i}\right| \leq O_p\left(a_n\right) \cdot \frac{1}{n} \sum_{i=1}^{n} \left|\overline{w}_i\right| = O_p\left(a_n\right) = o_p\left(1\right)$$

and thus

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \overline{w}_i \tilde{h}_i^{'} =& \mathbb{E}\left[\overline{w}_i \overline{h}_i^{'} \right] + \frac{1}{N} \sum_{i=1}^{N} \overline{w}_i \left(\tilde{h}_i - h_i\right)^{'} + \frac{1}{N} \sum_{i=1}^{N} \left(\overline{w}_i h_i^{'} - \mathbb{E}\left[\overline{w}_i h_i^{'} \right] \right) \\
=& \mathbb{E}\left[\overline{w}_i \overline{h}_i^{'} \right] + O_p\left(a_N\right) + O_p\left(\frac{1}{\sqrt{N}}\right) \xrightarrow{p} \Sigma_{wh} := \mathbb{E}\left[\overline{w}_i \overline{h}_i^{'} \right].
\end{aligned}$$

Hence,

$$\sqrt{N}\left(\hat{\alpha}-\alpha\right)=\left(\frac{1}{N}\sum_{i=1}^{N}\overline{w}_i\tilde{h}_i\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}g\left(z_i,\hat{\gamma}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\Sigma_{wh}^{-1}\Omega\Sigma_{wh}^{'-1}\right).$$

□

B.3 Proof of Theorem 2(ii)

Proof. We adapt the proof of Theorem 2(i) above with

$$\begin{aligned} g^*\left(z,\hat{\alpha},\hat{\gamma}\right):=&\overline{w}_i\left(\hat{\gamma}_j\left(h_{ji}\right)-\hat{\alpha}_0-\hat{\alpha}_jh_{ji}-\hat{\alpha}_k\hat{\gamma}_k^{-1}\left(\hat{\gamma}_j\left(h_{ji}\right)\right)\right),\\ g^*\left(z,\hat{\gamma}\right):=&\overline{w}_i\left(\hat{\gamma}_j\left(h_{ji}\right)-\alpha_0-\alpha_jh_{ji}-\alpha_k\hat{\gamma}_k^{-1}\left(\hat{\gamma}_j\left(h_{ji}\right)\right)\right).\end{aligned}$$

with $\mathbb{E}\left[g^*\left(z_i,\gamma\right)\right]=\mathbb{E}\left[\overline{w}_i\left(\gamma_j\left(h_{ji}\right)-\alpha_0-\alpha_jh_{ji}-\alpha_k\gamma_k^{-1}\left(\gamma_j\left(h_{ji}\right)\right)\right)\right]=\mathbb{E}\left[\overline{w}_iu_i\right]=\mathbf{0}$ and $\frac{1}{N}\sum_{i=1}^{N}g\left(z,\hat{\alpha}^*,\hat{\gamma}\right)=\mathbf{0}$.

By the chain rule,

$$\begin{aligned} G^*\left(z_i,\tau\right):=&\nabla_\gamma g^*\left(z,\gamma\right)\left[\hat{\gamma}-\gamma\right]\\ =&\overline{w}_i\left(\left[\hat{\gamma}_j\left(h_{ji}\right)-\gamma_j\left(h_{ji}\right)\right]-\alpha_k\nabla_\gamma\left(\gamma_k^{-1}\circ\gamma_j\right)\left[\hat{\gamma}-\gamma\right]\right)\\ =&\overline{w}_i\left(1-\frac{\alpha_k}{\gamma_k^{'}\left(h_{ki}\right)}\right)\left[\hat{\gamma}_j\left(h_{ji}\right)-\gamma_j\left(h_{ji}\right)\right]-\overline{w}_i\frac{\alpha_k}{\gamma_k^{'}\left(h_{ki}\right)}\left[\hat{\gamma}_k\left(h_{ki}\right)-\gamma_k\left(h_{ki}\right)\right]\\ =&\overline{w}_i\left[d_{i1}\left(1-\frac{\alpha_2}{\gamma_2^{'}},-\frac{\alpha_2}{\gamma_2^{'}}\right)+d_{i2}\left(-\frac{\alpha_1}{\gamma_1^{'}},1-\frac{\alpha_1}{\gamma_1^{'}}\right)\right]\left(\hat{\gamma}-\gamma\right)\end{aligned}$$

and

$$\int G\left(z_i,\hat{\gamma}-\gamma\right)\mathbb{P}z_i=\int\overline{w}_i\left(\lambda_1\left(1-\frac{\alpha_2}{\gamma_2^{'}}\right)+\lambda_2\frac{\alpha_1}{\gamma_1^{'}},\ \lambda_1\frac{\alpha_2}{\gamma_2^{'}}+\lambda_2\left(1-\frac{\alpha_1}{\gamma_1^{'}}\right)\right)\left(\hat{\gamma}-\gamma\right)d\mathbb{P}x_i$$

By Proposition 4 of Newey (1994), with

$$\varphi^*\left(z_i\right):=-\left(\lambda_1\left(1-\frac{\alpha_2}{\gamma_2^{'}}\right)+\lambda_2\frac{\alpha_1}{\gamma_1^{'}}\right)d_{i1}+\left(\lambda_1\frac{\alpha_2}{\gamma_2^{'}}+\lambda_2\left(1-\frac{\alpha_1}{\gamma_1^{'}}\right)\right)d_{i2}$$

we have

$$\overline{w}_i\left(\lambda_1\left(1-\frac{\alpha_2}{\gamma_2^{'}}\right)+\lambda_2\frac{\alpha_1}{\gamma_1^{'}},\ \lambda_1\frac{\alpha_2}{\gamma_2^{'}}+\lambda_2\left(1-\frac{\alpha_1}{\gamma_1^{'}}\right)\right)\begin{pmatrix} d_{i1}\left(q_i-\gamma_1\left(h_{1i}\right)\right)\\ d_{i2}\left(q_i-\gamma_2\left(h_{2i}\right)\right)\end{pmatrix}\equiv\varphi^*\left(z_i\right)\overline{w}_i\epsilon_i,$$

and by Assumption 8

$$\int G\left(z,\hat{\gamma}-\gamma\right)d\mathbb{P}\left(z\right)=\frac{1}{N}\sum_{i=1}^{N}\varphi^*\left(z_i\right)\overline{w}_i\epsilon_i+o_p\left(\frac{1}{\sqrt{N}}\right).$$

Hence, we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g^{*}\left(z_i,\hat{\gamma}\right)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[g^{*}\left(z_i,\gamma\right)+\varphi^{*}\left(z_i\right)\overline{w}_i\right]+o_p\left(1\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\Omega^{*}\right),$$

where

$$\Omega:=\mathrm{Var}\left[g^{*}\left(z_i,\gamma\right)+\delta^{*}\left(z_i\right)\right]=\mathbb{E}\left[\overline{w}_i\overline{w}_i^{'}\left(u_i^2+\varphi^{*}\left(z_i\right)^2\epsilon_i^2\right)\right],$$

giving

$$\sqrt{N}\left(\hat{\alpha}-\alpha\right)=\left(\frac{1}{N}\sum_{i=1}^{N}\overline{w}_i\tilde{h}_i\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}g^{*}\left(z_i,\hat{\gamma}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\Sigma_{wh}^{-1}\Omega^{*}\Sigma_{wh}^{'-1}\right).$$

□

B.4 Proof of Theorem 2(iii)

Proof. By (5), we have

$$\frac{\partial}{\partial c}\gamma_j\left(c;w\right)=\alpha_j+\alpha_k h_k^{'}\frac{1}{h_j^{'}}+\frac{1}{h_j^{'}}>\alpha_j,$$

and thus $0<\alpha_j/\gamma_j^{'}<1$, which implies

$$\lambda_1\left(1-\frac{\alpha_2}{\gamma_2^{'}}\right)+\lambda_2\frac{\alpha_1}{\gamma_1^{'}}>0,\quad \lambda_2\left(1-\frac{\alpha_1}{\gamma_1^{'}}\right)+\lambda_1\frac{\alpha_2}{\gamma_2^{'}}>0.$$

Hence,

$$\begin{aligned}\varphi^{*}&=\left(\lambda_1\left(1-\frac{\alpha_2}{\gamma_2^{'}}\right)+\lambda_2\frac{\alpha_1}{\gamma_1^{'}}\right)d_{i1}+\left(\lambda_2\left(1-\frac{\alpha_1}{\gamma_1^{'}}\right)+\lambda_1\frac{\alpha_2}{\gamma_2^{'}}\right)d_{i2}>0\\ 1+\varphi&=1-\left(\frac{\alpha_2}{\gamma_2^{'}}\lambda_1-\frac{\alpha_1}{\gamma_1^{'}}\lambda_2\right)\left(d_{i1}-d_{i2}\right)\\ &=\left(1-\lambda_1\frac{\alpha_2}{\gamma_2^{'}}+\lambda_2\frac{\alpha_1}{\gamma_1^{'}}\right)d_{i1}+\left(1-\lambda_2\frac{\alpha_1}{\gamma_1^{'}}+\lambda_1\frac{\alpha_2}{\gamma_2^{'}}\right)d_{i2}\\ &=\varphi^{*}+\left(1-\lambda_1\right)d_{i1}+\left(1-\lambda_2\right)d_{i2}\\ &>\varphi^{*}>0.\end{aligned}$$

Hence, $\left(1+\varphi\right)^2>\varphi^{*2}>0$ and

$$\Omega-\Omega^{*}=\mathbb{E}\left[\overline{w}_i\overline{w}_i^{'}\left[\left(1-\varphi\left(x_i,d_i\right)\right)^2-\varphi^{*}\left(x_i,d_i\right)^2\right]\epsilon_i^2\right]$$

is positive definite. □

B.5 Proof of Propositions 1 and 2

Proof. Assumption 9(i) guarantees that $N_1 \sim N_2 \sim N$ so that

$$\|\hat{\gamma}_1 - \gamma_1\|_\infty \sim \|\hat{\gamma}_2 - \gamma_2\|_\infty = O_p(a_N)$$

where, by Assumption 9(ii)-(v) and Theorem 8 of Hansen (2008),

$$a_N = b^p + \frac{\sqrt{\log N}}{\sqrt{Nb^3}}.$$

With b chosen according to Assumption 9(vi) so that $\frac{\sqrt{\log N}}{\sqrt{Nb^3}} = o\left(N^{-\frac{1}{4}}\right)$ and $\sqrt{N}b^p \to 0$, implying that

$$a_N = o\left(N^{-\frac{1}{2}}\right) + o\left(N^{-\frac{1}{4}}\right) = o\left(N^{-\frac{1}{4}}\right),$$

verifying Assumption 7(ii). Assumption 8 (and consequently Proposition 1) follows from Theorem 8.11 of Newey and McFadden (1994).

Since $\hat{\varphi} \xrightarrow{p} \varphi$ and $\hat{\varphi}^* \xrightarrow{p} \varphi^*$, Proposition 2 then follows from Theorem 8.13 of Newey and McFadden (1994). □