Using Monotonicity Restrictions to Identify Models with Partially Latent Covariates*

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

January 14, 2021

Abstract

This paper develops a new method for identifying econometric models with partially latent covariates. Such data structures arise naturally in industrial organization and labor economics settings where data are collected using an "input-based sampling" strategy, e.g., if the sampling unit is one of multiple labor input factors. We show that the latent covariates can be nonparametrically identified, if they are functions of a common shock satisfying some plausible monotonicity assumptions. With the latent covariates identified, semiparametric estimation of the outcome equation proceeds within a standard IV framework that accounts for the endogeneity of the covariates. We illustrate the usefulness of our method using two applications. The first focuses on pharmacies: we find that production function differences between chains and independent pharmacies may partially explain the observed transformation of the industry structure. Our second application investigates education achievement 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 Xu Cheng, Aureo de Paula, Ulrich Doraszelski, Amit Gandhi, Claudia Goldin, Aviv Nevo, Dan Silverman, Petra Todd, and seminar participants at numerous universities for comments and suggestions. Postlewaite and Sieg acknowledge support from the National Science Foundation.

1 Introduction

This paper develops a new method for identifying econometric models with partially latent covariates. We show that a broad class of econometric models that play a large role in industrial organization and labor economics can be nonparametrically identified if the partially latent covariate variables satisfy certain monotonicity assumptions. Examples that fall into this class of models are a variety of different production, skill formation, and achievement functions.[1] It is often plausible to assume that the different inputs or explanatory variables are functions of a common unobserved random shock, and we consider models in which it is natural to impose strict monotonicity in this common shock.[2] The monotonicity assumption imposes some strong functional dependencies on the explanatory variables as pointed out in the context of production function estimation by Ackerberg, Caves, and Frazer (2015). The key insight of this paper is that we can leverage the functional dependence between inputs to achieve identification within a partially latent covariate framework. In that sense, we turn the functional dependence problem on its head to impute the partially latent covariates. Broadly speaking, our imputation is in the spirit of matching algorithms (Rubin, 1973). In contrast to traditional matching algorithms, we propose to match on the expected dependent variable to impute missing covariates.[3]

The partially latent data structure, that we study in this paper, arises quite naturally in many potential applications of our technique if one employs an "input-based sampling" strategy, i.e. if the sampling unit is one of the multiple labor input factors. These types of data sets are becoming more prevalent in modern econometrics since researchers have come to rely on unstructured or semi-structured data sets. Consider, for example, a production team in which team members perform different tasks. Let us assume that the researcher interviews one member from each team to provide the data. It is plausible that this person knows the team's output, but does not have

[1] Other potential applications in applied microeconomics are discussed in the conclusions.

[2] Note that this assumption is commonly used, for example, in the production function literature as discussed by Olley and Pakes (1996). In particular, this assumption does not require that inputs are "optimally" chosen by competitive firms and is consistent with a broad class of strategic and non-strategic models that may describe the agents' behavior.

[3] Note that we do not apply the matching approach within the standard potential outcome framework of program evaluation which is based on the potential outcome model developed by Fisher (1935). For a discussion of the properties of matching estimators in that context see, among others, Rosenbaum and Rubin (1983), Heckman, Ichimura, Smith, and Todd (1998), and Abadie and Imbens (2006).

complete information about the other team members' input choices. By randomly sampling the teams we elicit information from all different types of team members and hence input factors. We call this type of sampling an "input-based sampling" approach and provide a formal definition of this data structure. Alternatively, consider a child that is raised by divorced or separated parents. It is likely that either one of the parents knows his or her inputs as well as the child's achievement but does not perfectly know the level of input provided by the divorced partner. Again, we show that we can identify the input functions of the father and the mother under the input-based sampling approach if inputs satisfy a plausible monotonicity restriction.

Once we have identified the latent covariates, the estimation of the outcome function can proceed using standard semiparametric methods developed in the econometric literature. One key issue here is that the common shock creates an endogeneity problem.[4] We show that we can combine our identification results with a variety of linear, nonlinear, and semiparametric estimation strategies. In that sense our approach is flexible and allows researchers to make appropriate functional form assumptions if necessary. To illustrate the key issues that are encountered in estimation we consider the scenario in which researchers only have access to a single cross-section of data and rely on instrumental variables for estimation.[5] For example, production function estimation relies on the assumption that differences in local input prices give rise to differences in input choices that are uncorrelated with productivity shocks at the local level.[6] Similarly, skill formation and achievement function estimation requires the choice of suitable instruments for parental inputs.[7]

Estimation proceeds in two steps. In finite samples, we first nonparametrically estimate the latent input functions. Plugging the estimators into our production,

[4]In the context of production function estimation this endogeneity problem is referred to as the transmission bias problem since inputs are correlated with unobserved productivity shocks (Marschak and Andrews, 1944).

[5]Hence we cannot address this endogeneity problem using panel data with fixed effects, first advocated by Hoch (1955, 1962) and Mundlak (1961, 1963). We can also not use more sophisticated timing assumptions within a control function or IV frameworks 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). We discuss the extension of our methods to this scenario in the conclusions.

[6]Hence, local input prices can serve as valid instruments for endogenous input choices. See Griliches and Mairesse (1998) for a critical discussion of the assumption that these input prices are exogenous.

[7]For a more general discussion of the issues encountered in estimating achievement and skill formation functions see, among others, Todd and Wolpin (2003) and Cunha, Heckman, and Schennach (2010).

skill formation, or achievement function, we can estimate the parameters of this outcome function using a standard IV estimator based on the observed and imputed covariates. The second econometric challenge then arises for the need to account for the sequential nature of the estimator when deriving the correct rate of convergence and computing asymptotic standard errors. To illustrate this we consider the standard log-linear, Cobb-Douglas model. We propose two different estimators and provide both high-level and lower-level conditions under which these semiparametric two-step estimators are consistent and asymptotically normal at the usual parametric rate of convergence. The technical proofs are 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). Finally, we show that using the conditional expectation of outcomes as the dependent variable produces efficiency gains relative to the more traditional estimator that uses the observed output instead.

To evaluate the performance of our estimator we conduct a variety of Monte Carlo experiments. Our findings suggest that our estimators are 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 instruments are also partially latent.

We then illustrate the usefulness of the techniques developed in this paper and consider 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 that have replaced independent pharmacies in many local markets. Here we estimate a team production function that distinguishes between managerial and non-managerial certified pharmacists. We can, therefore, test the hypothesis whether managers have become more productive in chains than in independent pharmacies.

We use data from the National Pharmacist Workforce Survey in 2000 which uses an "input-based sampling" procedure. It not only collects data for each pharmacist that is surveyed but also a limited amount of information at the store level including output. We find that 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 skill formation or achievement functions which play a large role in pubic, labor, and family economics. Here we rely on data from the Child Development Supplement 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 latent variable IV estimator produces similar results to the feasible IV estimator. We also consider a sample of children from divorced households where the father's inputs have to be imputed. Hence, the standard IV estimator is no longer feasible, but our latent variable IV estimator can still be applied. We find that there are some significant differences between married and divorced parents. In particular, divorced fathers have no significant impact on child quality.

This paper relates 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 leveraged by Olley and Pakes (1996) and Levinsohn and Petrin (2003). They 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. While most of these papers focus on value-added production functions, there is also much interest in estimating gross output production functions. Doraszelski and Jaumandreu (2013) propose an solution to the transmission bias problem that also relies on observed firm-level variation in prices. In particular, they show that by explicitly imposing the parameter restrictions between the production function and the demand for a flexible input and by using this price variation, they can recover the gross output production function. Gandhi, Navarro, and Rivers (2020) provide an alternative identification strategy to estimate gross output production functions that works well in short panels. Beyond these conceptual linkages, our paper has a different focus from these papers cited above: they focus more on the dynamic nature of capital inputs, while we focus on the problem of partially latent inputs. Moreover, the estimation of production functions is just one of many applications of our general identification result. This paper shows that our methods may be even more useful for applications outside of IO where these data structures are more prevalent as we discuss below.

Also, 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 in a scalar unobservable shock to identify and impute the latent input.

Similarly, our monotonicity conditions also differentiate our paper from the econometric literature on data combination as surveyed by Ridder and Moffitt (2007), which mostly involves conditional independence assumptions. That said, in a way our proposed method can be regarded as a strategy to combine two samples, each of which contains a common outcome variable and a different covariate variable. Hence, our proposed method may also be useful as a data combination method for scenarios where our monotonicity conditions are interpretable and justifiable.

The rest of the paper is organized as follows. Section 2 presents our main identification result. Section 3 discusses the problems associated with estimation. 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. Section 6 provides a discussion of other applications and presents our conclusions.

2 Identification of Partially Latent Covariates

2.1 Model and Main Result

Consider the following cross-sectional econometric model

$$y_i = F(x_{i1}, x_{i2}, u_i) + \epsilon_i \tag{1}$$

where $i = 1, ..., N$ indexes a generic observation from a *random sample*, y_i denotes an observable scalar-valued outcome variable, and $x_i := (x_{i1}, x_{i2})$ denotes a two-dimensional vector of covariates.[8] Both u_i and ϵ_i are scalar-valued unobserved errors, with u_i taken to be a "structural error" that is endogenous with respect to x_i, while ϵ_i is a "measurement error" that is assumed to be exogenous. The unknown outcome function F may be either parametric or nonparametric.

First, we need to define what we mean by *partially latent covariates*, a key data structure that we explore in this paper.

Assumption 1 (Partially Latent Covariates)**.** *For each observation i, the econometrician either observes x_{i1} or x_{i2}, but never both.*

Essentially, one of the two covariates (x_{i1}, x_{i2}) is latent in each observation in the data. In the following, it will be convenient to write

$$d_i := \begin{cases} 1, & \text{if } x_{i1} \text{ is observed and } x_{i2} \text{ is latent,} \\ 2, & \text{if } x_{i2} \text{ is observed and } x_{i1} \text{ is latent,} \end{cases}$$

so that effectively $(d_i, (2 - d_i) x_{i1}, (d_i - 1) x_{i2})$ is observed for i. Such data structures often arise when the data is collected at the individual level while we are interested in some firm, household, or team level outcome variable that also depends on other individuals who are not surveyed in the data. These types of unstructured data sets are becoming increasingly more prevalent in empirical work, as we discuss in detail below. In this section we just provide one application that we use as the leading example to illustrate the main concepts.

Example (Team Production Functions)**.** Our first application studied in Section 4

[8] See Corollary 1 for the extension of our identification method to settings with covariates of higher dimensions.

focuses on identifying and estimating team production functions.[9] For simplicity, let us assume a log-linear Cobb-Douglas specification:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + u_i + \epsilon_i, \tag{2}$$

where y_i is the logarithm of the team's output, x_{i1} is the logarithm of hours worked by the first team member (a manager), and x_{i2} is the logarithm of hours worked by the second team member (an employee).[10] The data structure described in Assumption 1 arises if the researcher interviews only one member, and not both members of the team. We also refer to this technique as an "*input-based sampling*" approach. It is plausible that the interviewed team member knows the team's output, but does not have complete information about the other team member's input choices. Hence, the surveyed person provides the output level, y_i, and her own hours worked, x_{i1} *or* x_{i2}, leading to the problem of partially latent inputs as defined in Assumption 1.

The next assumption imposes a monotonicity condition on the outcome function.

Assumption 2 (Monotonicity of the Outcome Function)**.** *F is nondecreasing in all of its arguments and is strictly increasing in at least one of its arguments.*

This assumption essentially states that the covariates or inputs (x_{i1}, x_{i2}) and the structural error u_i have nonnegative effects on the outcome variable y_i. Moreover, the monotonicity is strict in, at least, one of the three arguments x_{i1}, x_{i2}, and u_i. The restriction of monotonicity with respect to (x_{i1}, x_{i2}) is substantive: it requires that the covariates cannot negatively affect the outcome variable holding everything else fixed. In contrast, the restriction of monotonicity with respect to u_i is largely innocuous given the interpretation of u_i as a (weakly) "positive shock".

Example (Team Production Functions Continued)**.** Assumption 2 is satisfied in the linear additive model in equation (2) provided that the model satisfies the additional parameter restriction that $\alpha_1, \alpha_2 \geq 0$.

[9] We use the term "team production function" since we largely focus on different types of labor inputs and abstract from capital or other inputs that may be subject to dynamics and adjustment costs.

[10] The team production concept is also related to the concept of task production functions, which are surveyed by Acemoglu and Autor (2011). Haanwinckel (2018) estimates a task production function in which each team member specializes in a single task.

Next, we turn to the assumptions on the unobserved errors u_i and ϵ_i in equation (1). First, we assume that the endogenous covariates x_i are strictly monotone functions of the scalar structural error u_i, potentially after conditioning on a set of observed covariates z_i, that may affect the covariates x_i.

Assumption 3 (Strict Monotonicity of the Covariates in the Structural Error). *There exists a vector of additional observed covariates z_i and two deterministic, real-valued functions h_1, h_2, such that*

$$x_{i1} = h_1(u_i, z_i), \quad x_{i2} = h_2(u_i, z_i),$$

with both $h_1(u_i, z_i)$ and $h_2(u_i, z_i)$ strictly increasing in their first argument u_i for every realization of z_i.

We note that the functions h_1 and h_2 can be unknown and nonparametric. Moreover, Assumption 3 does not require z_i to be exogenous; in other words, z_i and u_i are allowed to be statistically dependent. The only requirement here is that, after conditioning on z_i, the covariates x_{i1} and x_{i2} can be written as deterministic monotone functions of the error u_i. Such a "monotonicity-in-a-scalar-error" assumption has been widely used in the econometric literature on identification analysis.[11]

Example (Team Production Functions Continued). In the IO literature u_i is typically interpreted as a "productivity shock" that enters into the choices of inputs x_i. In contrast, ϵ_i captures either a measurement error or a productivity shock that does not affect inputs, since it is not observed to the firms when input choices are made. Assumption 3 requires that the input choice functions are strictly increasing in the "productivity shock" u_i, conditional on any additional observed covariates z_i that may influence input choices, as suggested, for example, by Olley and Pakes (1996) and others.[12] For concreteness, we take z_i to be local wages for managers and employees.

The monotonicity of input choices in the unobserved productivity shock can be further micro-founded in a variety of settings based on efficiency or equilibrium criteria. For example, Assumption 3 is automatically satisfied if competitive firms optimally

[11] See Matzkin (2007) for a general survey, and see Ackerberg, Caves, and Frazer (2015) in the specific context of production function identification, which fits into our working example (2).

[12] This 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.

choose inputs to maximize profits. The input choice functions 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_i and decreasing in z_i.[13] More generally, one may use the theory of monotone comparative statics to obtain more primitive conditions for input monotonicity, which typically involve various forms of increasing-difference or single-crossing conditions: see, for example, Milgrom and Shannon (1994) and Vives (2000) for formal statements. Essentially, in settings where input choices are made by a single decision maker, such as under perfect competition and monopsony, we would need the marginal values of inputs to be increasing in the productivity shock u_i, which is a mild condition to impose given our interpretation of u_i as a "productivity shock". In settings where the input choices are generated as equilibria of a strategic game between two decision makers, an additional assumption of strategic complementarity is typically sufficient for monotonicity. For games with strategic substitutability, we would further need a condition to ensure that the extent of strategic substitutability is not overwhelming: see Roy and Sabarwal (2010) for general results, and our Appendix D for an example where Assumption 3 is satisfied under strategic substitutability.

Next we formalize the required exogeneity condition on the measurement error ϵ_i.

Assumption 4 (Exogeneity of the Measurement Error)**.** $\mathbb{E}\left[\epsilon_i \middle| x_i, z_i, d_i\right] = 0$.

Note that, under Assumption 3, conditioning on (x_i, z_i, d_i) is equivalent to conditioning on (u_i, z_i, d_i). In the production function estimation literature without the partial latency problem, $\mathbb{E}\left[\epsilon_i \middle| u_i, z_i\right] = 0$ is a standard assumption imposed on ϵ_i. In our current setting, we are requiring that ϵ_i is furthermore exogenous with respect to the partial latency indicator variable d_i.

It is worth noting 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

[13] See Appendix A for details. 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). However, 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.

usually exploits two types of assumptions to handle missing values: first, observations with no missing data occur with positive probability, and second, data are "missing at random (MAR)": the indicator for missingness is exogenous to or independent of certain observable covariates or constructed conditioning variables. Neither condition is satisfied in our setting: here every observation contains "missing values", and the partial latency indicator d_i is allowed to be correlated with other observables as well as the unobserved productivity shock. Instead, we will be relying on monotonicity conditions to identify and impute the latent input.

Specifically, Assumption 4 here is simply requiring that ϵ_i is a "measurement error" term that is exogenous with respect to the observables and consequently the "productivity shock" u_i, but does not impose any restriction on the dependence structure between the partial latency indicator d_i and other structural components of the model (u_i, x_i, z_i).

However, we do require the following very mild condition on the variable d_i.

Assumption 5 (Nondegenerate Latency Probabilities)**.** $0 < \mathbb{P}\{d_i = 1|\, u_i, z_i\} < 1$.

Assumption 5 guarantees that conditioning on realizations of (u_i, z_i) we will observe x_{i1}, and x_{i2}, with strict positive probabilities. Again, this assumption is much weaker than "missing-at-random" assumptions, which would usually require that $\mathbb{P}\{d_i = 1|\, u_i, z_i\}$ is constant in u_i, z_i, or some other variables. In contrast, here we do not impose any restrictions on the dependence of $\mathbb{P}\{d_i = 1|\, u_i, z_i\}$ on (u_i, z_i) beyond non-degeneracy.

We are now ready to present our main identification result.

Theorem 1. *Under Assumptions 1-5, for each observation i, the latent covariate, x_{i2} if $d_i = 1$ or x_{i1} if $d_i = 2$, is point identified.*

Next, we provide a detailed explanation of our identification strategy. The starting point of our identification strategy is the reduced form of our model with the measurement error term:

$$y_i = \overline{F}\left(u_i, z_i\right) + \epsilon_i \tag{3}$$

where

$$\overline{F}\left(u_i, z_i\right) := F\left(h_1\left(u_i, z_i\right), h_2\left(u_i, z_i\right), u_i\right). \tag{4}$$

Clearly, $\overline{F}(u_i, z_i)$ is strictly increasing in u_i given Assumptions 2 and 3.

Consider two firms i and j with $z_i = z_j$. In the context of our working example, we are effectively considering two firms i and j operating in the same local labor market with the same local wages. For concreteness, suppose that (x_{i1}, x_{j1}) are observed, while (x_{i2}, x_{j2}) are unobserved. Since these firms have the same value of managerial inputs $x_{i1} = x_{j1}$, then by Assumption 3 it must also be true that they have the same value of the productivity shock:

$$u_i = h_1^{-1}(x_{i1};\ z_i) = h_1^{-1}(x_{j1};\ z_j) = u_j,$$

where $h_1^{-1}(\cdot\,; z_i)$ is the inverse of $h_1(\cdot, z_i)$, which is well-defined by Assumption 3. This further implies that

$$\overline{F}(u_i, z_i) = \overline{F}(u_j, z_j).$$

Taking an average of y_i and y_j,

$$\frac{1}{2}(y_i + y_j) = \overline{F}(u_i, z_i) + \frac{1}{2}(\epsilon_i + \epsilon_j), \tag{5}$$

we are essentially averaging out the variations in ϵ.[14] Intuitively, if we average over outcomes of all observations that share the same x_{i1} and the same z_i and thus the same value of u_i, then we can identify $\overline{F}(u_i, z_i)$.

Formally, define $\gamma_1(c)$ as the expected output of firm i conditional on the event that x_{i1} is observed $(d_i = 1)$ to have a given value of c_1, i.e.,

$$\gamma_1(c_1; z) := \mathbb{E}\left[y_i \mid z_i = z,\ d_i = 1,\ x_{i1} = c_1\right]. \tag{6}$$

Clearly, γ_1 is directly identified from data given Assumptions 1 and 5,[15] and can be nonparametrically estimated later on. Taking a closer look at γ_1, we have, by equation (3), Assumption 3, and Assumption 4,

$$\begin{aligned}\gamma_1(c_1; z) &= \mathbb{E}\left[\overline{F}(u_i, z_i) + \epsilon_i \middle| z_i = z, d_i = 1, h_1(u_i, z_i) = c_1\right] \\ &= \overline{F}\left(h_1^{-1}(c_1; z), z\right) + \mathbb{E}\left[\epsilon_i \middle| z_i = z, d_i = 1, u_i = h_1^{-1}(c_1; z)\right] \\ &= F\left(c_1,\ h_2\left(h_1^{-1}(c_1; z), z\right),\ h_1^{-1}(c_1; z)\right),\end{aligned} \tag{7}$$

[14] In fact, we can directly "match" on output y_i if there is no measurement error, ϵ_i, in output.

[15] Assumption 5 ensures that the conditioning event occurs with strictly positive probability.

which is a direct formalization of the intuition in equation (5). By conditioning on z_i and a particular *observed value* of $x_{i1} = c_1$, we are effectively conditioning on the *unobserved* productivity shock u_i. Aggregating across observations allows us to average out the measurement errors and obtain a quantity that is implicitly a function of the productivity shock $u_i = h_1^{-1}(c_1;\ z_i)$.

Next, we observe that $\gamma_1(c_1; z)$ is strictly increasing in c_1, since

$$\frac{\partial}{\partial c_1}\gamma_1(c_1; z) = F_1 + F_2 \cdot \frac{\partial}{\partial u} h_2\left(h_1^{-1}(c_1), z\right) \frac{1}{\frac{\partial}{\partial u} h_1\left(h_1^{-1}(c_1), z\right)} + F_3 \cdot \frac{1}{\frac{\partial}{\partial u} h_1\left(h_1^{-1}(c_1), z\right)} > 0 \tag{8}$$

since $\frac{\partial}{\partial u}h_1, \frac{\partial}{\partial u}h_2 > 0$ by Assumption 3, and the partial derivatives F_1, F_2, F_3 of F are all nonnegative with, at least, one being strictly positive by Assumption 2.[16] Similarly, we can define

$$\gamma_2(c_2; z) := \mathbb{E}\left[y_i | z_i = z,\ d_i = 2,\ x_{i2} = c_2\right]$$

which is strictly increasing in c_2.

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

$$\gamma_1(c_1; z) = \gamma_2(c_2; z) \tag{9}$$

for some c_1, c_2, and z.

Example (Team Production Functions Continued)**.** Let us consider production teams within the same local market so that wages (z_i) are constant. Equation (9) then involves two separate conditional expected output levels, one (γ_1) for teams whose manager input (x_{i1}) is observed , and the other (γ_2) for teams whose employee input (x_{i2}) is observed. When these two expected output levels are equalized as in equation (9), we can infer that the underlying productivity shock (u_i) must be the same across all teams with either $x_{i1} = c_1$ observed or $x_{i2} = c_2$ observed. By equations (5) and (7) we know

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

[16] The partial derivatives F_1, F_2, F_3 of F are evaluated at $\left(c_1,\ h_2\left(h_1^{-1}(c_1; z), z_i\right),\ h_1^{-1}(c_1; z)\right)$.

which also pins down the latent inputs via:

$$x_{i2} = h_2(\overline{u}, z_i), \quad \text{for } d_i = 1,$$
$$x_{i1} = h_1(\overline{u}, z_i), \quad \text{for } d_i = 2.$$

Formally, the latent covariates can be identified via a composition of γ_1, γ_2 and their inverses,

$$\begin{aligned} x_{i2} &= \gamma_2^{-1}(\gamma_1(x_{i1}; z_i); z_i), \quad \text{for } d_i = 1, \\ x_{i1} &= \gamma_1^{-1}(\gamma_2(x_{i2}; z_i); z_i), \quad \text{for } d_i = 2, \end{aligned} \tag{10}$$

since on the right-hand side x_{i1}, x_{i2} are observed for $d_i = 1, 2$, respectively, and γ_1, γ_2 are nonparametrically identified functions. This completes the description of our key identification strategy as well as the proof of Theorem 1.

Remark 1 (More Than Two Covariates). We have thus far focused on the case with two covariates. It is straightforward to see that our model, assumptions, and the main identification result can be easily generalized to the case with covariates of an arbitrary finite dimension D. This result is summarized by the following Corollary.

Corollary 1. *Consider the model $y_i := F(x_{i1}, ..., x_{iD}, u_i) + \epsilon_i$ along with Assumptions 2 and 4 unchanged, and the following modifications of other assumptions:*

(i) Assumption 1: for each i at least one out of D covariates is observed.

(ii) Assumption 3: all D covariates are strictly increasing in u_i given z_i.

(iii) Assumption 5: all D covariates are observed with strictly positive probabilities.

Then the latent covariates are identified.

Remark 2. If Condition (i) in Corollary 1 is strengthened so that *more than one* covariates are simultaneously observed in a given observation (with positive probability), then we would also obtain over-identification, and the input-monotonicity restriction in Assumption 3 becomes empirically refutable. Alternatively, with two or more covariates simultaneously observed, we would be able to accommodate higher dimensions of unobserved shocks, provided that the dimension of the unobserved shock u_i is strictly smaller than the dimension of the covariates D. Since such an extension would be more involved and move farther away from the applications we consider in this paper, we leave it as a direction for future research.

3 Identification and Estimation of Outcome Functions

With the latent inputs already identified in Theorem 1, we are back to equation (1)

$$y_i = F(x_{i1}, x_{i2}, u_i) + \epsilon_i,$$

but now we can effectively regard both x_{i1} and x_{i2} as being known, at least for identification purposes. Researchers may proceed to identify the output function F under appropriate application-specific assumptions as in a "standard" setting without the partial latency problem.

Hence, the identification of F or other objects of interest is largely "separable" from the partial latency problem, which is the key problem we are solving in this paper. That said, we note that the *estimation* of the latent covariates will affect the *estimation* of (the parameters of) F based on "plugged-in" latent covariate estimates. This section provides a discussion on how to identify and estimate F, and analyzes the impact of the "first-stage" estimation of latent inputs on the final estimator of F.

While we cannot cover all relevant specifications of F, in this section we will provide both identification and estimation results for the linear case, which is arguably the workhorse model, or at least a natural benchmark, in various empirical applications. We also discuss how our method can be applied under more general settings.

3.1 The Linear Model

In this subsection we focus on the linear parametric specification of F as in (2):

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + u_i + \epsilon_i,$$

where our goal is to identify and estimate the unknown parameters $\alpha := (\alpha_0, \alpha_1, \alpha_2)$.

3.1.1 Identification

In the presence of the endogeneity problem between $x_i := (x_{i1}, x_{i2})$ and u_i, we will need instrumental variables for the identification of α. For illustrational simplicity, we impose the following standard IV assumption.

Assumption 6 (Instrumental Variables)**.** *Write* $z_i := (z_{i1}, z_{i2})$, $\overline{z}_i := (1, z_{i1}, z_{i2})'$ *and* $\overline{x}_i = (1, x_{i1}, x_{i2})'$. *Assume*

(i) Relevance: $\Sigma_{zx} := \mathbb{E}\left[\overline{z}_i \overline{x}_i'\right]$ *has full rank.*

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

Corollary 2 (Identification of Linear Parameters)**.** *Under Assumptions* (1)-(6)*,* α *is point identified.*

Example (Team Production Function Continued)**.** In the context of our working example, here we are essentially following a strategy discussed in Griliches and Mairesse (1998) and assume that we have access to some instrumental variables (such as local wages) that affect input choices.

3.1.2 Estimation Procedure

We now turn to the more interesting problem of estimation, propose semiparametric estimators for α, and characterize their asymptotic distributions.

We first describe our proposed estimator. Since the identification of latent inputs via equation (10) is constructive, it suggests a natural estimation procedure:

Step 1 (Nonparametric Regression): obtain an estimator $\hat{\gamma}_1$ of γ_1 by nonparametrically regressing y_i on x_{i1} and z_i, among firms with $d_i = 1$, i.e., those with x_{i1} observed. 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 (10), i.e.,

$$\begin{aligned} \hat{x}_{i2} &= \hat{\gamma}_2^{-1}\left(\hat{\gamma}_1\left(x_{i1}; z_i\right); z_i\right), \quad \text{for } d_i = 1, \\ \hat{x}_{i1} &= \hat{\gamma}_1^{-1}\left(\hat{\gamma}_2\left(x_{i2}; z_i\right); z_i\right), \quad \text{for } d_i = 2. \end{aligned}$$

Step 3 (IV Regression): estimate equation (2) with z_i as IVs for x_i, i.e.,

$$\hat{\alpha} := \left(\frac{1}{n}\sum_{i=1}^{n}\overline{z}_i\tilde{x}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\overline{z}_i y_i\right)$$

and

$$\tilde{x}_i := \begin{cases}(1, x_{i1}, \hat{x}_{i2})', & \text{for } d_i = 1,\\ (1, \hat{x}_{i1}, x_{i2})', & \text{for } d_i = 2.\end{cases}$$

In Appendix B.4, we also propose an alternative estimator $\hat{\alpha}^*$ that features a slightly different Step 3, leading to an efficiency gain over $\hat{\alpha}$ asymptotically. Since the asymptotic theories for $\hat{\alpha}$ and $\hat{\alpha}^*$ are very similar, we defer results on $\hat{\alpha}^*$ to the appendix.

3.1.3 Asymptotic Theory

We now establish the consistency and the asymptotic normality of $\hat{\alpha}$ under the following regularity assumptions.

Assumption 7 (Finite Error Variances)**.** $\mathbb{E}\left[u_i^2\middle| z_i\right] < \infty$ *and* $\mathbb{E}\left[\epsilon_i^2\middle| x_i, z_i, d_i\right] < \infty$.

Assumption 8 (Strong Monotonicity)**.** *The first derivative of* $\gamma_k(\cdot, z)$ *is uniformly bounded away from zero, i.e., for any* c, z*,*

$$\frac{\partial}{\partial c}\gamma_k(c; z) > \underline{c} > 0.$$

In view of equation (8), Assumption 8 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 8 is needed to ensure that $\hat{\gamma}_k^{-1}(\cdot, z)$ is a good estimator of $\gamma_k^{-1}(\cdot, z)$ provided that the first-stage nonparametric estimator $\hat{\gamma}_k$ is consistent for γ_k.

Assumption 9 (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 9(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. Assumption 9(ii) requires that the first-stage estimator converges at a rate faster than $N^{-1/4}$, which is satisfied under various types of nonparametric estimators under certain regularity conditions. This 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 $\hat{\alpha}$ through the functional derivatives of the residual function with respect to γ_1, γ_2. Assumption 10 below, based on Newey (1994), provides an explicit formula for the asymptotic variance of $\hat{\alpha}$ that does not depend on the particular forms of first-stage nonparametric estimators.

Formally, write $w_i := (y_i, x_i, z_i, d_i)$, $\gamma := (\gamma_1, \gamma_2)$, and suppress the conditioning variables z_i in γ for notational simplicity. Define the residual functions

$$g\left(w_i, \tilde{\alpha}, \tilde{\gamma}\right) := \begin{cases} \overline{z}_i \left(y_i - \tilde{\alpha}_0 - \tilde{\alpha}_1 x_{i1} - \tilde{\alpha}_2 \tilde{\gamma}_2^{-1}\left(\tilde{\gamma}_1\left(x_{i1}\right)\right)\right) & \text{for } d_i = 1, \\ \overline{z}_i \left(y_i - \tilde{\alpha}_0 - \tilde{\alpha}_2 x_{i2} - \tilde{\alpha}_1 \tilde{\gamma}_1^{-1}\left(\tilde{\gamma}_2\left(x_{i2}\right)\right)\right) & \text{for } d_i = 2. \end{cases}$$

for generic $\tilde{\alpha}, \tilde{\gamma}$, and $g\left(w_i, \tilde{\gamma}\right) := g\left(w_i, \alpha, \tilde{\gamma}\right)$ at the true α. Define the pathwise functional derivative of g at γ along direction τ by

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

Then, following Newey (1994), the so-called "influence function" can be derived analytically[17] based on G and takes the form of $\varphi\left(w_i\right) \overline{z}_i \epsilon_i$ with

$$\varphi\left(w_i\right) := -\left(\lambda_1 \frac{\alpha_2}{\gamma_2^{\prime}} - \lambda_2 \frac{\alpha_1}{\gamma_1^{\prime}}\right)\left(\mathbb{1}\left\{d_i = 1\right\} - \mathbb{1}\left\{d_i = 2\right\}\right),$$

where $\gamma_k^{\prime}$ denotes $\frac{\partial}{\partial h_k} \gamma_k\left(x_{ik}; z_i\right)$, λ_1 stands for

$$\lambda_1\left(x_i, z_i\right) := \mathbb{E}\left[\mathbb{1}\left\{d_i = 1\right\} \middle| x_i, z_i\right]$$

i.e., the conditional probability of observing x_{i1}, and $\lambda_2 := 1 - \lambda_1$.

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

The influence function essentially characterizes how the first-stage estimation influences the asymptotic variance of the final estimator. Formally, we present the following assumption, commonly known as an asymptotic linearity condition, which basically requires that the expected error induced by the first-stage estimation is asymptotically equivalent to the sample average of $\varphi(w_i)\,\overline{z}_i\epsilon_i$. In particular, the formula for φ given above will be the same regardless of the specific forms of first-step estimators used, provided that some suitable regularity conditions are satisfied.

Assumption 10 (Asymptotic linearity)**.** *Suppose*

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

We emphasize that Assumptions 9 and 10 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 11 below, we also provide an example of lower-level conditions that replace Assumptions 9 and 10 when we use the Nadaraya-Watson kernel estimator in the first-stage nonparametric regression.

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

Theorem 2 (Asymptotic Normality)**.** *Under Assumptions 1-10,*

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

where $\Sigma := \Sigma_{zx}^{-1}\Omega\Sigma_{xz}^{-1}$ *and*

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

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 $\Sigma_{zx}^{-1}\mathrm{Var}\left(\overline{z}_i\left(u_i + \epsilon_i\right)\right)\Sigma_{xz}^{-1}$. Hence, the presence of the additional term $\delta(z_i)$ in Ω captures the effect of the first-step nonparametric regression on the asymptotic variance of $\hat{\alpha}$.

To obtain consistent variance estimators, define

$$\hat{\Omega} := \frac{1}{N}\sum_{i=1}^{N}\overline{z}_i\overline{z}_i'\left[y_i - \tilde{x}_i'\hat{\alpha} + \hat{\varphi}(w_i)(y_i - \tilde{y}_i)\right]^2$$

where

$$\tilde{y}_i := \begin{cases}\hat{\gamma}_1(x_{i1}, z_i), & \text{for } d_i = 1,\\ \hat{\gamma}_2(x_{i2}, z_i), & \text{for } d_i = 2,\end{cases}$$

and with

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

where $\hat{\lambda}_1$ is any consistent nonparametric estimator of λ_1. Then the variance estimators can be obtained as

$$\hat{\Sigma} := S_{x\tilde{z}}^{-1}\hat{\Omega}S_{\tilde{z}x}^{-1}$$

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

Proposition 1. *In addition to Assumptions 1-8 and 11, suppose that $\hat{\lambda}_1$ is any consistent nonparametric estimator of λ_1. Then $\hat{\Omega} \xrightarrow{p} \Omega$ and $\hat{\Omega}^* \xrightarrow{p} \Omega^*$.*

If furthermore $\lambda_1(x_i, z_i) \equiv \lambda_1 \in (0, 1)$ is assumed, then we may use the sample proportion $\hat{\lambda}_1 := \frac{1}{N}\sum_i\{d_i = 1\}$.

3.1.4 Lower-Level Regularity Conditions for Kernel First Step

Finally, we present a set of lower-level conditions that replace Assumptions 9 and 10, when we use the canonical Nadaraya-Watson kernel estimator for the nonparametric regression in Step 1. We emphasize that this subsection simply serves as an illustration of Assumptions 9-10 and Theorem 2, as our method does not require the use of a 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 11 (Example of Lower-Level Conditions with Kernel First Step)**.** *Let $N_k := \sum_{i=1}^{N}\mathbb{1}\{d_i = k\}$ denote the number of firms for which h_{ik} is observed, and let*

$\hat{\gamma}_k$ be the Nadaraya-Watson kernel estimator of γ_k defined by

$$\hat{\gamma}_k(v) := \frac{\frac{1}{N_k b^3}\sum_{d_i=k} K\left(\frac{v-v_{ik}}{b^3}\right) y_i}{\frac{1}{N_k b^3}\sum_{d_i=k} K\left(\frac{v-v_{ik}}{b^3}\right)}$$

where $v_{ik} := (x_{ik}, z_{i1}, z_{i2})$ for all i such that $d_i = k$. Suppose the following conditions:

(i) $\lambda_1(x_i, z_i) \in (\epsilon, 1-\epsilon)$ *for all* (x_i, z_i) *for some* $\epsilon > 0$.

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

(iii) $\mathbb{E}\left[y_i^4\right] < \infty$ *and* $\mathbb{E}\left[y_i^4 \middle| x_i, z_i\right] f(x_i, z_i)$ *is bounded.*

(iv) γ_k *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 11(i) essentially requires that the proportion of observations with x_{i1} observed and that with x_{i2} 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 x_{i1} and γ_2 based on observations with x_{i2} well enough asymptotically. Assumption 11(iv) is the key smoothness condition that will help establish the Donsker property (and a consequent stochastic equicontinuity condition) in Assumption 9(i). Assumption 11(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 "under-smoothed" way) so that the kernel estimator $\hat{\gamma}_k$ converges at a rate faster than $N^{-1/4}$, as required in Assumption 9(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 9-10 are satisfied. See Newey and McFadden (1994, Section 8.3) for additional details.

Proposition 2 (Asymptotic Distributions with Kernel First Step)**.** *Under Assumptions 1-8 and 11, the conclusions of Theorem 2 hold.*

3.2 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

$$\Sigma := \left(\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}\right)^{-1}\Sigma_{xz}\Sigma_{zz}^{-1}\Omega\Sigma_{zz}^{-1}\Sigma_{zx}\left(\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}\right)^{-1}.$$

Other Parametric Outcome Function

Consider a potentially nonlinear parametric production function of the form

$$y_i = F_\alpha\left(x_{i1}, x_{i2}\right) + u_i + \epsilon_i$$

After the identification of partially latent inputs via Theorem 1, the second stage boils down to the estimation of α based on the moment condition $\mathbb{E}\left[z_i\left(y_i - F_\alpha\left(x_{i1}, x_{i2}\right)\right)\right] = \mathbf{0}$, which can be obtained via GMM estimation. Technically, since GMM estimators are Z-estimators, the corresponding asymptotic theory in Newey and McFadden (1994), on which the proof of Theorem 2 is mainly based, still applies with proper changes in notation.

Nonparametric Outcome Function

More generally, with any nonparametric production function that is additively separable in u_i and ϵ_i of the form

$$y_i = F\left(x_{i1}, x_{i2}\right) + u_i + \epsilon_i,$$

where F is an unknown function that satisfies Assumption 2, the only thing that changes is the second-stage nonparametric estimation of F with the imputed covariates $\tilde{x}_i$ (or more precisely, with one component known and one component imputed) based on the moment condition $\mathbb{E}\left[z_i\left(y_i - F\left(x_{i1}, x_{i2}\right)\right)\right] = \mathbf{0}$. The asymptotic theory for this case can be similarly obtained based on theory on nonparametric two-step estimation (e.g. Ai and Chen, 2007, and Hahn, Liao, and Ridder, 2018).

In the more general specification (1):

$$y_i = F(x_{i1}, x_{i2}, u_i) + \epsilon_i$$

where there is no more additive separability in u_i, one way to obtain identification and implement IV estimation is by adapting Chernozhukov, Imbens, and Newey (2007) to our current context. Essentially, we would need to impose strict monotonicity of F in u_i, impose independence of u_i from z_i, normalize the distribution of u_i to be uniform, and then exploit a quantile-based residual condition as described in Chernozhukov, Imbens, and Newey (2007).

3.3 A Monte Carlo Experiment

Here we report the findings of some Monte Carlo experiments. Table 1 reports the parameter specifications of the Cobb-Douglas production function that we use in our experiments. 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 of 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 x_{i1} and x_{i2} have a linear relationship with y_i. We estimate $\gamma_1(\cdot, z_i)$ and $\gamma_2(\cdot, z_i)$ using second degree polynomials. Not surprisingly, we find that the estimated coefficients on quadratic terms are almost 0. The interpolated functions γ_1^{-1} and γ_2^{-1} are also almost linear.

Table 2 summarizes the performance of two different estimators: TSLS when all inputs are observed as well as our version of TSLS when inputs are imputed. We refer to our version of the TSLS estimator as the "matched" TSLS estimator . 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

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	$\begin{pmatrix}0.5 & 0\\0 & 0.5\end{pmatrix}$

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 variable using within market information. Instead, we pool observations across markets and estimate conditional expectations conditional on x_1 (or x_2), z_1, and z_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:

$$(z_{i1}, z_{i2}) = \begin{cases} (z^*_{i1}, missing) & \text{if } x_{i1} \text{ is observed} \\ (missing, z^*_{i2}) & \text{if } x_{i2} \text{ is observed} \end{cases} \tag{11}$$

Since we need to impute missing wages, we assume that true wages are functions of some demand shifters $D_m \in \mathbb{R}^2$ 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

Table 2: Monte Carlo: Different Markets, Observed Wages

Param	Number of Markets	Number of Firms	Spec	TSLS Bias	TSLS RMSE	Matched TSLS Bias	Matched TSLS 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

specification allows for correlation between $z_{1m(i)}$ and $z_{2m(i)}$ through D_m. Specifically, we simulate wages as follows:

$$
\begin{aligned}
z_{i1}^* &= z_{1m(i)} = \kappa_1 D_{1m} + \kappa_2 D_{2m} + \eta_{i1} \\
z_{i2}^* &= z_{2m(i)} = \kappa_3 D_{1m} + \kappa_4 D_{2m} + \eta_{i2}
\end{aligned}
\tag{12}
$$

To impute the missing wages, we regress the observed wages (z_{i1}, z_{i2}) on the demand shifters (D_{1m}, D_{2m}). Using estimated parameters from the regression, we then impute the missing wages.

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

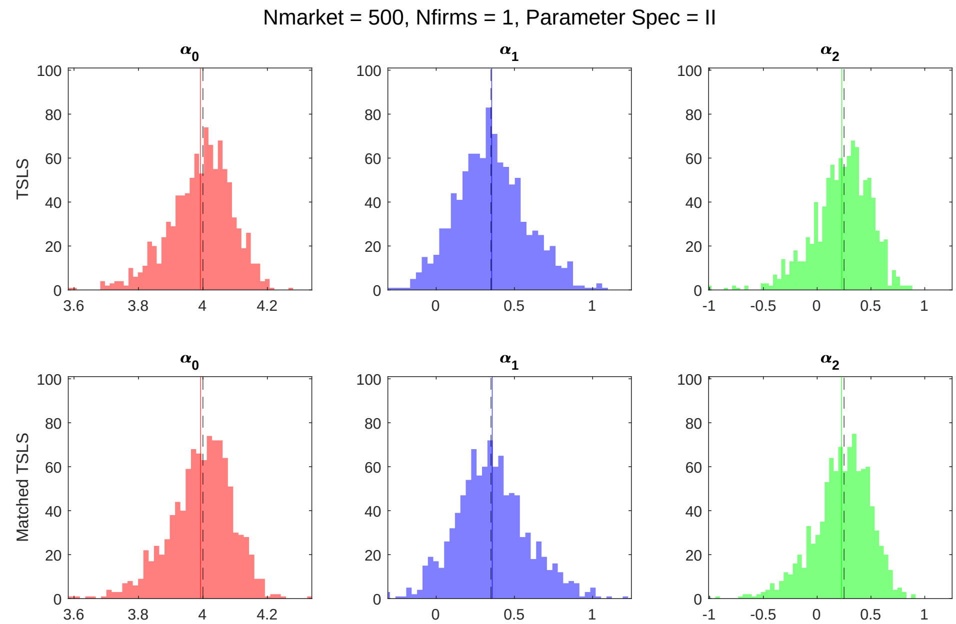
	Number of	Number of		Standard SLS		Matched TSLS	
Param	markets	firms	Spec	Bias	RMSE	Bias	RMSE
α_0	500	1	1	-0.004	0.003	-0.004	0.003
α_0	500	1	2	-0.008	0.010	-0.007	0.010
α_0	500	1	3	-0.008	0.010	-0.007	0.010
α_1	500	1	1	-0.002	0.015	-0.001	0.016
α_1	500	1	2	0.005	0.054	0.008	0.055
α_1	500	1	3	0.004	0.053	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 more general than those considered in Sections 3 of the paper. In particular, we do not need to observe both sets of instruments in the data, i.e. 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

Figure 1: Histograms of Estimated Coefficients With Imputed Wages



production functions.

4 First Application: Pharmacies

Our first application focuses on the industrial organization of pharmacies. This industry has undergone a dramatic change over the past 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 types 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, the quality of the firm-level data is likely to be 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 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.[18] We explore these issues in more detail below and test whether the different types of pharmacies have access to the same technology.

The survey 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 the employees 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.[19] Hence, we need to use the version of our estimator that averages across local markets as discussed in Section 3.3.

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 (16) in Appendix A. Under the assumption of Cobb-Douglas production, we can test the optimality by jointly testing the null hypothesis of equality of both coefficients. Table 6 shows the results. A formal Wald test rejects the null hypothesis of optimality. Thus the direct inversion of the optimality conditions cannot be applied to estimate the parameters of the production function, whereas our new estimator is feasible.

We implement two versions of our "matched" TSLS estimator: the first estimator uses the observed outputs while the second one uses expected outputs. Since the observed output is subject to a measurement error, the semi-parametric estimator

[18]Most pharmacies in our sample have one manager pharmacist and one employee pharmacist, but there are a few pharmacies with a larger employment size. See Appendix C for details on how to compute employees' hours work for the pharmacies with multiple employees.

[19]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.

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 parentheses.

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)

using expected outputs offers the potential of some efficiency gains as discussed in Appendix B.4. 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 for weak instruments. Overall, we find that our instruments are sufficiently strong in most cases.[20]

Table 7 shows that we estimate most of 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.[21]

Table 7: Estimation Result

	Independent		Chain	
	Observed Outputs	Expected Outputs	Observed Outputs	Expected Outputs
α_0	5.447	5.857	2.504	3.634
	(0.597)	(0.331)	(1.790)	(1.060)
α_1	0.227	0.163	0.819	0.687
	(0.122)	(0.057)	(0.454)	(0.268)
α_2	0.090	0.047	0.409	0.250
	(0.071)	(0.051)	(0.191)	(0.105)
Nobs	144	144	188	188
First-stage F for x_1	9.320	9.320	11.774	11.774
First-stage F for x_2	13.648	13.648	3.630	3.630

Our results provide several insights to understanding the difference between in-

[20]As a robustness check, we also explored a different matching algorithm which estimates the expectation of output conditional on local demand shifters rather than wages. The results are consistent although the matching algorithm with local demand shifters gives slightly larger point estimates with slightly less precision.

[21]Appendix C provides some additional robustness checks.

dependents and chains. First, our results indicate that chains may have a different production function than independent pharmacies. A formal joint hypothesis test reported in Table 8 rejects the null hypothesis that the coefficients of the production function are the same.

Table 8: Hypothesis Tests

	Production Function (Joint)	Managerial Efficiency α_1	Residual Variance $V(u)$
Independent		0.163	0.010
Chain		0.687	0.006
Difference or Ratio		-0.524	1.532
Test Statistics	122.841	-1.913	1.532
Test	*Wald*	t	F
p-value	(0.000)	(0.028)	(0.003)

Second, our findings also suggest that managers may be more effective in chains than independents. A formal one-sided t-test reported in Table 8 rejects the null hypothesis that the two coefficients that characterize managerial efficiency are the same.

Finally, we find that chains have a significantly lower residual variance than independents. A formal F test reported in Table 8 rejects the null hypothesis that the residual variance of independents is greater than or equal to the residual variance of chains. Note that all the tests are based on the estimation results with the expected outputs as the dependent variable.

We thus conclude that chains have different production functions than independent pharmacies which may partially explain the change in the observed market structure of that industry. However, more research is needed to fully address this important research question.

5 Second Application: Child Education

Our second application focuses on the estimation of education achievement functions. Here we assume that a child's achievement y_i is a function of the mother's and the father's time inputs, denoted by x_{im} and x_{if}. Again, we consider a log-linear Cobb-

Douglas specification given by

$$y_i \;=\; \alpha_i \,+\, \alpha_m\, x_{im} \,+\, \alpha_f\, x_{if} \,+\, u_i \tag{13}$$

where heterogeneity in the intercept is given by:

$$\alpha_i \;=\; x_i'\, \alpha_0 \tag{14}$$

Hence, we assume that the 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 partially latent.

Our data is based on the four available waves of the Child Development Supplement (CDS). These are the cohorts interviewed in 1997, 2002, 2007, and 2014.[22] 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.[23] The CDS 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.[24]

We begin by estimating an education production function using the subsample of children who live in married households. Hence, we observe the mother's and the father's inputs in the data set. We observe 3,236 children with complete inputs and applied problem scores as well as 2,789 children with complete inputs and reading

[22] 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.

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

[24] 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.

comprehension scores. Table 9 provides descriptive statistics of the main variables in our sample.

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

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. 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 measured in

10,000 dollars. Table 10 summarizes our findings.

Table 10: Education Production Function: Married Sample

	Applied Problems		Passage Comprehension	
	TSLS	matched TSLS	TSLS	matched TSLS
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)
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)
Constant	4.510	4.484	4.321	4.380
	(0.017)	(0.026)	(0.026)	(0.223)
Nobs	3,236	3,236	2,789	2,789
First-stage F for x_m	61.997	127.295	41.812	58.530
First-stage F for x_f	62.636	117.966	58.654	59.156

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 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 all households in this sample 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 in-

Table 11: Education Production Function: Divorced Sample

	Applied Problems matched TSLS	Passage Comprehension matched TSLS
Mom Hour	0.050	0.037
	(0.028)	(0.015)
Dad Hour	0.010	0.001
	(0.013)	(0.003)
Num Child = 2	0.051	0.019
	(0.055)	(0.039)
Num Child = 3+	0.002	−0.015
	(0.056)	(0.066)
Household Labor Inc	−0.013	−0.006
	(0.016)	(0.004)
Constant	4.548	4.529
	(0.078)	(0.061)
Nobs	785	723
First-stage F for x_m	40.532	35.264
First-stage F for x_f	15.715	56.184

put, which we use for imputation purposes.[25] Note that the standard TSLS is no longer feasible in this subsample because of the latent variable problem. Table 11 summarizes our findings.

Table 11 shows that the time inputs for mothers are positive, statistically significant, and economically meaningful. Moreover, the point estimates for the applied problem test are similar to the ones we obtained for the married sample reported in Table 10. The main difference is that mother's time inputs are slightly less productive for children from divorced families, and father's time inputs are not statistically different from zero. In summary, our estimator work well in this application and yields plausible and accurate point estimates for most coefficients of interest. Most importantly, we find that the inputs of divorced fathers into the skill formation function of their children seem to be negligible.

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

6 Concluding Remarks

We have developed a new method for identifying econometric models with partially latent covariates. We have shown that a broad class of econometric models that play a large role in industrial organization and labor economics can be non-parametrically identified if the partially latent covariates are monotonic functions of a common shock. Examples that fall into this class of models are production and skill formation functions. The partially latent data structure arises quite naturally in these settings if we employ an "input-based sampling" strategy, i.e. if the sampling unit is one of multiple labor input factors. It is plausible that the sampling unit will only have incomplete information about the other labor inputs that affect output. Our proofs of identification are constructive and imply a sequential, two-step semi-parametric estimation strategy. We have discussed the key problems encountered in estimation, characterized rate of convergence, and the asymptotic distribution of our estimators.

We also presented two applications of our technique. Our first application focuses on estimating team production functions. Using a national survey of pharmacists, we have found some convincing evidence that chains have different technologies than independently operated pharmacies. In particular, managers appear to be more productive in chains. Our second application focuses on the estimation of skill formation functions, which play a large role in labor and family economics. We have shown that our matched TSLS estimator produces similar results to the feasible 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 must be imputed. We find that the inputs of divorced fathers into the skill formation function of their children is negligible.

There is substantial scope for future research in areas other than the two applications that we provided above. At the heart of the applications discussed thus far is the relationship between multiple inputs that are combined to produce a single output. It is easy to imagine questions that ask about relationships that fit this structure and that do not fall into the frameworks we have considered thus far.

To illustrate this idea, consider the problem of inter vivos gifts. It is common for parents, while still alive, to give money to their children, often to help with a down payment on a house or to reduce taxes the parents will pay. When a couple makes a gift to their married child, however, they risk that the child divorces and a portion of the gift will accrue to the child's spouse. The concern is real since approximately 40% of marriages in the US end in divorce. A natural question is how well can parents predict how long a child's marriage will last at the time they contemplate making a gift. One could address this question with a data set that includes inter vivos gifts from parents to married children and, in addition, how long the child's marriage survives. Such data sets exist, for example the PSID, which documents these for a family lines that stretch over a half century.

There is a problem however: Multigenerational data sets such as PSID have quite detailed information about the choices of individuals who are descendants of the initial respondents, but substantially less information about choices of individuals who "marry into" the data set. For each married couple in the PSID, one of the two has the "PSID gene" (that is, a descendant of an initial respondent), and we have substantially more information about that individual and, importantly, about that individual's parents than we have about the spouse. In particular, we know the inter vivos gifts to the couple from the parents of the PSID gene child but not inter vivos gifts to the couple from the spouse's parents. Note that this design of the PSID gives rise to a data structure that mimics the "input-based sampling" approach that we have studied in this paper.[26] As we show in Appendix D, it is straightforward to write down a non-cooperative model of intergenerational transfer, where the transfers of each parents are monotonically increasing in the probability that the marriage survives. This potential application is an example of interesting problems that arise in trying to understand intergenerational effects. We would like to know how the choices or characteristics of individuals in one generation affect the outcomes of their descendants. We conjecture that the methods developed in this paper can be fruitfully applied to study a variety of questions related to intergenerational linkages.

Finally, our research provides ample score for future research in econometric methodology. We have restricted ourselves to applications in which our method of identification can be combined with standard IV techniques to estimate the functions of interest. Much of the recent panel data literature has focused on dynamic inputs in the presence of adjustment costs. More research is clearly needed to evaluate whether the ideas presented in this paper can be extended and applied to dynamic panel data frameworks. We have also restricted ourselves to systems of inputs with a single com-

[26] Other multigenerational data sets such as NLSY79, NLSY97 and NCDS share the partially latent variable problem.

mon shock. Another potentially interesting research question is how our methods can be extended to more complicated econometric structures with multiple shocks.

References

Abadie, A. and G. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74 (1), 235–67.

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.

Acemoglu, D. and D. Autor (2011): "Skills, Tasks and Technologies: Implications for Employment and Earnings," in *Handbook of Labor Economics*, ed. by D. Card and O. Ashenfelter, Elsevier, 1043–1171.

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.

Ai, C. and X. Chen (2007): "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables," *Journal of Econometrics*, 141, 5–43.

Bergstrom, T., L. Blume, and H. Varian (1986): "On the Private Provision of Public Goods," *Journal of Public Economics*, 29, 25–49.

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.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, 139, 4–14.

Cunha, F., J. Heckman, and S. Schennach (2010): "Estimating the Technology of Cognitive and Non-cognitive Skill Formation." *Econometrica*, 78, 883–931.

Doraszelski, U. and J. Jaumandreu (2013): "R & D and Productivity: Estimating Endogenous Productivity," *Review of Economic Studies*, 80, 1338–83.

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

Fisher, R. (1935): *Design of Experiments*, Hafner, new York.

Gandhi, A., S. Navarro, and D. Rivers (2020): "On the Identification of Gross Output Production Functions," *Journal of Political Economy*, 128, 2973–3016.

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.

Haanwinckel, D. (2018): "Supply, Demand, Institutions, and Firms: A Theory of Labor Market Sorting and the Wage Distribution," Working Paper.

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.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias using Experimental Data," *Econometrica*, 66 (2), 315–331.

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.

Matzkin, R. L. (2007): "Nonparametric identification," *Handbook of econometrics*, 6, 5307–5368.

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

Milgrom, P. and C. Shannon (1994): "Monotone comparative statics," *Econometrica: Journal of the Econometric Society*, 157–180.

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.

Ridder, G. and R. Moffitt (2007): "The Econometrics of Data Combination," *Handbook of econometrics*, 6, 5469–5547.

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.

Rosenbaum, P. and D. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrica*, 70, 41–55.

Roy, S. and T. Sabarwal (2010): "Monotone comparative statics for games with strategic substitutes," *Journal of Mathematical Economics*, 46, 793–806.

Rubin, D. (1973): "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183.

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

Todd, P. and K. Wolpin (2003): "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal*, 113, F3–33.

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

Vives, X. (2000): *Oligopoly pricing: Old ideas and new trends*, Cambridge, MA: MIT Press.

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:

$$\max_{X_{i1},X_{i2}} e^{\alpha_0+u_i} X_{i1}^{\alpha_1} X_{i2}^{\alpha_2}\ e^{u_i} - Z_{i1}X_{i1} - Z_{i2}X_{i2}, \tag{15}$$

where $X_{i1}, X_{i2}, Z_{i1}, Z_{i2}$ denote exponents of $x_{i1}, x_{i2}, z_{i1}, z_{i2}$. By the first-order conditions,

$$\begin{aligned}
X_{i1} &= e^{\frac{\alpha_0+u_i}{1-\alpha_1-\alpha_2}} \left(\frac{Z_{i1}}{\alpha_1}\right)^{\frac{1-\alpha_2}{\alpha_1+\alpha_2-1}} \left(\frac{Z_{i2}}{\alpha_2}\right)^{\frac{\alpha_2}{\alpha_1+\alpha_2-1}} \\
X_{i2} &= e^{\frac{\alpha_0+u_i}{1-\alpha_1-\alpha_2}} \left(\frac{Z_{i2}}{\alpha_2}\right)^{\frac{1-\alpha_1}{\alpha_1+\alpha_2-1}} \left(\frac{Z_{i1}}{\alpha_1}\right)^{\frac{\alpha_1}{\alpha_1+\alpha_2-1}} \\
\overline{Y}_i &= e^{\frac{\alpha_0+u_i}{1-\alpha_1-\alpha_2}} \left(\frac{Z_{i1}}{\alpha_1}\right)^{\frac{\alpha_1}{\alpha_1+\alpha_2-1}} \left(\frac{Z_{i2}}{\alpha_2}\right)^{\frac{\alpha_2}{\alpha_1+\alpha_2-1}} \\
&= e^{\alpha_0+u_i} \left(\frac{\alpha_2 Z_{i1}}{\alpha_1 Z_{i2}}\right)^{\alpha_2} x_{i1}^{\alpha_1+\alpha_2} = e^{\alpha_0+u_i} \left(\frac{\alpha_1 Z_{i2}}{\alpha_2 Z_{i1}}\right)^{\alpha_1} x_{i2}^{\alpha_1+\alpha_2}
\end{aligned}$$

In log forms

$$\begin{aligned}
x_{i1} = h_1(u_i, z_i) &= \frac{\alpha_0 + (1-\alpha_2)\log\alpha_1 + \alpha_2\log\alpha_2}{1-\alpha_1-\alpha_2} - \frac{1-\alpha_2}{1-\alpha_1-\alpha_2} z_{i1} - \frac{\alpha_2}{1-\alpha_1-\alpha_2} z_{i2} + \frac{1}{1-\alpha_1-\alpha_2} u_i \\
x_{i2} = h_2(u_i, z_i) &= \frac{\alpha_0 + \alpha_1\log\alpha_1 + (1-\alpha_1)\log\alpha_2}{1-\alpha_1-\alpha_2} - \frac{\alpha_1}{1-\alpha_1-\alpha_2} z_{i1} - \frac{1-\alpha_1}{1-\alpha_1-\alpha_2} z_{i2} + \frac{1}{1-\alpha_1-\alpha_2} u_i \\
\overline{y}_i = \overline{y}(u_i, z_i) &= \frac{\alpha_0 + \alpha_1\log\alpha_1 + \alpha_2\log\alpha_2}{1-\alpha_1-\alpha_2} - \frac{\alpha_1}{1-\alpha_1-\alpha_2} z_{i1} - \frac{\alpha_2}{1-\alpha_1-\alpha_2} z_{i2} + \frac{1}{1-\alpha_1-\alpha_2} u_i \\
&= \alpha_0 + \alpha_2\log(\alpha_2/\alpha_1) + (\alpha_1+\alpha_2) h_1(u_i, z_i) + \alpha_2 z_{i1} - \alpha_2 z_{2l} + u_i \\
&= \alpha_0 + \alpha_1\log(\alpha_1/\alpha_2) + (\alpha_1+\alpha_2) h_2(u_i, z_i) - \alpha_1 z_{i1} + \alpha_1 z_{2l} + u_i
\end{aligned}$$

Taking inverses

$$\begin{aligned}
u_i = h_1^{-1}(x_{i1}, z_i) &:= -\left[\alpha_0 + (1-\alpha_2)\log\alpha_1 + \alpha_2\log\alpha_2\right] + (1-\alpha_1-\alpha_2) x_{i1} + (1-\alpha_2) z_{i1} + \alpha_2 z_{i2} \\
= h_2^{-1}(x_{i2}, z_i) &:= -\left[\alpha_0 + \alpha_1\log\alpha_1 + (1-\alpha_1)\log\alpha_2\right] + (1-\alpha_1-\alpha_2) x_{i2} + \alpha_1 z_{i1} + (1-\alpha_1) z_{i2}
\end{aligned}$$

Hence,

$$\gamma_1(x_{i1}, z_i) = \overline{y}\left(h_1^{-1}(x_{i1}, z_i), z_i\right) = -\log\alpha_1 + x_{i1} + z_{i1},$$
$$\gamma_2(x_{i2}, z_i) = \overline{y}\left(h_2^{-1}(x_{i2}, z_i), z_i\right) = -\log\alpha_2 + x_{i2} + z_{i2},$$

and

$$\begin{aligned} y_i &= \gamma_1(x_{i1}, z_i) + \epsilon_i = -\log\alpha_1 + x_{i1} + z_{i1} + \epsilon_i \\ &= \gamma_2(x_{i2}, z_i) + \epsilon_i = -\log\alpha_2 + x_{i2} + z_{i2} + \epsilon_i. \end{aligned} \tag{16}$$

It is then evident that α_1 or α_2 can be estimated directly from (16) from the corresponding subsample where x_{i1} or x_{i2} is observed. Furthermore, we may test input optimality based on equation (16).

B Proofs

B.1 Additional Notation and Lemmas

Notation For each i, we use x_{ij} to denote the observed input and use x_{ik} to denote the latent input variable for firm i, i.e.

$$x_{ij} = x_{i1},\ x_{ik} = x_{i2},\ \text{for } d_i = 1,$$
$$x_{ij} = x_{i2},\ x_{ik} = x_{i1},\ \text{for } d_i = 2.$$

We write

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

so that $x_{ij} = d_{i1}x_{i1} + d_{i2}x_{i2}$ while $x_{ik} := d_{i1}x_{i2} + d_{i2}x_{i1}$. We write $\overline{x}_i := (1, x_{i1}, x_{i2})'$ to denote the true regressor vector. (Recall $\tilde{x}_i$ denotes the same regressor vector with imputed latent input $\hat{x}_{ik}$ in place of x_{ik}.)

Moreover, we suppress the instrumental variables z_i in functions, such as $\gamma_1(u_i, z_i)$, unless it becomes necessary to emphasize the dependence of such functions on z_i.

Lemma 1. *Under Assumption 8, 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* $|\hat{x}_{ik} - x_{ik}| = O_p(a_n)$.

Proof. By Assumption 8 we have

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

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

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

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 8 and Lemma 1, for $d_i = 1$,

$$\begin{aligned}|\hat{x}_{ik} - x_{ik}| &= \left|\hat{\gamma}_j^{-1}(\hat{\gamma}_k(x_{ik})) - \gamma_j^{-1}(\gamma_k(x_{ik}))\right| \\ &= \left|\hat{\gamma}_j^{-1}(\hat{\gamma}_k(x_{ik})) - \gamma_j^{-1}(\hat{\gamma}_k(x_{ik})) + \gamma_j^{-1}(\hat{\gamma}_k(x_{ik})) - \gamma_j^{-1}(\gamma_k(x_{ik}))\right| \\ &\leq \left|\hat{\gamma}_j^{-1}(\hat{\gamma}_k(x_{ik})) - \gamma_j^{-1}(\hat{\gamma}_k(x_{ik}))\right| + \left|\gamma_j^{-1}(\hat{\gamma}_k(x_{ik})) - \gamma_j^{-1}(\gamma_k(x_{ik}))\right| \\ &\leq \left\|\hat{\gamma}_j^{-1} - \gamma_j^{-1}\right\|_\infty + \frac{1}{\underline{c}}|\hat{\gamma}_k(x_{ik}) - \gamma_k(x_{ik})| \\ &\leq \left\|\hat{\gamma}_j^{-1} - \gamma_j^{-1}\right\|_\infty + \frac{1}{\underline{c}}\|\hat{\gamma}_k - \gamma_k\|_\infty \\ &= O_p(a_n). \end{aligned} \tag{17}$$

□

Lemma 2. *Under Assumption 8:*

(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

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

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

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

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 8 and γ'' and τ_k' are uniformly bounded above by Assumption 9(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,

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

□

B.2 Proof of Theorem 2

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

Recall $w_i := (y_i, x_i, z_i, d_i)$, $\gamma := (\gamma_1, \gamma_2)$ and

$$\begin{aligned}
g\left(w_i,\hat\alpha,\hat\gamma\right)=&\overline{z}_i\left(y_i-\hat\alpha_0-\left(x_{i1}\hat\alpha_1+\hat\gamma_2^{-1}\left(\hat\gamma_1\left(x_{i1}\right)\right)\hat\alpha_2\right)d_{i1}-\left(x_{i2}\hat\alpha_2+\hat\gamma_1^{-1}\left(\hat\gamma_2\left(x_{i2}\right)\right)\hat\alpha_2\right)d_{i2}\right)\\
=&\overline{z}_i\left(y_i-\hat\alpha_0-x_{ij}\hat\alpha_j-\hat\gamma_k^{-1}\left(\hat\gamma_j\left(x_{ij}\right)\right)\hat\alpha_k\right)\\
g\left(w_i,\hat\gamma\right)=&\overline{z}_i\left(y_i-\alpha_0-\left(x_{i1}\alpha_1+\hat\gamma_2^{-1}\left(\hat\gamma_1\left(x_{i1}\right)\right)\alpha_2\right)d_{i1}-\left(x_{i2}\alpha_2+\hat\gamma_1^{-1}\left(\hat\gamma_2\left(x_{i2}\right)\right)\alpha_2\right)d_{i2}\right)\\
=&\overline{z}_i\left(y_i-\alpha_0-x_{ij}\alpha_j-\hat\gamma_k^{-1}\left(\hat\gamma_j\left(x_{ij}\right)\right)\alpha_k\right)\\
=&\overline{z}_i\left(u_i+\epsilon_i+\left[x_{ik}-\hat\gamma_k^{-1}\left(\hat\gamma_j\left(x_{ij}\right)\right)\right]\alpha_k\right)
\end{aligned}$$

Clearly, $\mathbb{E}\left[g\left(w_i,\gamma\right)\right]=\mathbb{E}\left[\overline{z}_i\left(u_i+\epsilon_i\right)\right]=0$ by Assumptions 6 and 4. Moreover, $\frac{1}{N}\sum_{i=1}^N g\left(w_i,\hat\alpha,\hat\gamma\right)=0$ by the definition of $\hat\alpha$.

Now, define

$$
\begin{aligned}
G\left(w_i, \hat{\gamma}-\gamma\right) &:= \nabla_\gamma g\left(w_i, \gamma\right)[\hat{\gamma}-\gamma] \\
&= -\alpha_k \overline{z}_i \nabla_\gamma\left(\gamma_k^{-1} \circ \gamma_j\right)[\hat{\gamma}-\gamma] \\
&= \frac{-\alpha_k \overline{z}_i}{\gamma_k^{\prime}\left(\gamma_k^{-1}\left(\gamma_j\left(x_{ij}\right)\right)\right)}\left[\left(\hat{\gamma}_j-\gamma_j\right)\left(x_{ij}\right)-\left(\hat{\gamma}_k-\gamma_k\right)\left(\gamma_k^{-1}\left(\gamma_j\left(x_{ij}\right)\right)\right)\right] \\
&= -\frac{\alpha_k \overline{z}_i}{\gamma_k^{\prime}\left(x_{ik}\right)}\left[\hat{\gamma}_j\left(x_{ij}\right)-\gamma_j\left(x_{ij}\right)-\hat{\gamma}_k\left(x_{ik}\right)+\gamma_k\left(x_{ik}\right)\right] \text{ since } \gamma_k^{-1}\left(\gamma_j\left(x_{ij}\right)\right)=x_{ik} \\
&= d_{i1} \overline{z}_i\left(-\frac{\alpha_2}{\gamma_2^{\prime}}\right)(1,-1)\begin{pmatrix}\hat{\gamma}_1-\gamma_1 \\ \hat{\gamma}_2-\gamma_2\end{pmatrix}+d_{i2} \overline{z}_i\left(-\frac{\alpha_1}{\gamma_1^{\prime}}\right)(-1,1)\begin{pmatrix}\hat{\gamma}_1-\gamma_1 \\ \hat{\gamma}_2-\gamma_2\end{pmatrix} \\
&= -\overline{z}_i\left(d_{i1} \frac{\alpha_2}{\gamma_2^{\prime}}-d_{i2} \frac{\alpha_1}{\gamma_1^{\prime}}\right)(1,-1)(\hat{\gamma}-\gamma) \qquad (18)
\end{aligned}
$$

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

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

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

Next, the stochastic equicontinuity condition

$$
\frac{1}{\sqrt{N}} \sum_{i=1}^N\left(G\left(w_i, \hat{\gamma}-\gamma\right)-\int G\left(w_i, \hat{\gamma}-\gamma\right) d \mathbb{P}\left(w_i\right)\right)=o_p\left(\frac{1}{\sqrt{N}}\right) \qquad (19)
$$

is guaranteed by Assumptions 8 and 9. Specifically, $\hat{\gamma}-\gamma$ belongs to a Donsker class of functions by the smoothness assumption while $1/\gamma_k^{\prime}\left(x_{ik}\right) \leq 1/\underline{c}$ guarantees that $G\left(z_i, \cdot\right)$ is square-integrable, so that $G\left(z_i, \cdot\right)$ is also Donsker and thus (19) holds.

Now, write $\zeta_i:=\left(x_i, z_i\right)$ so that $w_i=\left(y_i, \zeta_i, d_i\right)$. Then we have

$$
\begin{aligned}
& \int G\left(w_i, \hat{\gamma}-\gamma\right) \mathbb{P} w_i \\
= & \int-\overline{z}_i\left(d_{i1} \frac{\alpha_2}{\gamma_2^{\prime}}-d_{i2} \frac{\alpha_1}{\gamma_1^{\prime}}\right)(1,-1)(\hat{\gamma}-\gamma) d \mathbb{P}\left(\zeta_i, d_i\right) \\
= & \int-\overline{z}_i\left(\left[\int d_{i1} d \mathbb{P}\left(d_i \mid \zeta_i\right)\right] \frac{\alpha_2}{\gamma_2^{\prime}}-\left[\int d_{i2} d \mathbb{P}\left(d_i \mid \zeta_i\right)\right] \frac{\alpha_1}{\gamma_1^{\prime}}\right)(1,-1)(\hat{\gamma}-\gamma) d \mathbb{P} \zeta_i \\
= & \int-\overline{z}_i\left(\lambda_1\left(\zeta_i\right) \frac{\alpha_2}{\gamma_2^{\prime}}-\lambda_2\left(\zeta_i\right) \frac{\alpha_1}{\gamma_1^{\prime}}\right)(1,-1)(\hat{\gamma}-\gamma) d \mathbb{P} \zeta_i
\end{aligned}
$$

By Proposition 4 of Newey (1994), with

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

we have

$$\overline{z}_i \left(\lambda_1 \frac{\alpha_2}{\gamma_2'} - \lambda_2 \frac{\alpha_1}{\gamma_1'}\right)(1, -1) \begin{pmatrix} d_{i1}(y_i - \gamma_1(x_{i1})) \\ d_{i2}(y_i - \gamma_2(x_{i2})) \end{pmatrix} \equiv \varphi(w_i)\, \overline{z}_i \epsilon_i,$$

and by Assumption 10

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

Hence, Lemma 5.4 of Newey (1994),

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

where

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

Lastly, by Lemma 1

$$\left|\frac{1}{n} \sum_{i=1}^{n} \overline{z}_i (\hat{x}_{i1} - x_{i1})\right| \leq \frac{1}{n} \sum_{i=1}^{n} |\overline{z}_i|\, |\hat{x}_{i1} - x_{i1}| \leq O_p(a_n) \cdot \frac{1}{n} \sum_{i=1}^{n} |\overline{z}_i| = O_p(a_n) = o_p(1)$$

and thus

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

Hence,

$$\sqrt{N}\left(\hat{\alpha}-\alpha\right)=\left(\frac{1}{N}\sum_{i=1}^{N}\overline{z}_i\tilde{x}_i\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}g\left(w_i,\hat{\gamma}\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\Sigma_{zx}^{-1}\Omega\Sigma_{zx}^{\prime -1}\right).$$

□

B.3 Proof of Propositions 2 and 1

Proof. Assumption 11(i) guarantees that $N_1 \sim N_2 \sim N$ so that

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

where, by Assumption 11(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 11(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 9(ii). Assumption 10 (and consequently Proposition 2) follows from Theorem 8.11 of Newey and McFadden (1994).

Since $\hat{\varphi} \xrightarrow{p} \varphi$ and $\hat{\varphi}^* \xrightarrow{p} \varphi^*$, Proposition 1 then follows from Theorem 8.13 of Newey and McFadden (1994). □

B.4 An Alternative and More Efficient Estimator $\hat{\alpha}^*$

The estimator $\hat{\alpha}$ proposed in the main text is defined by an IV estimator of the regression equation

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + u_i + \epsilon_i, \quad \mathbb{E}\left[u_i+\epsilon_i \middle| z_i\right]=0$$

in Step 3, where the left-hand side is the raw outcome variable y_i. Alternatively, with Steps 1 and 2 unchanged, we may construct a slightly different estimator $\hat{\alpha}^*$ for α based on the conditionally expected outcome as described below.

Step 3*: Estimate the following equation

$$\overline{y}_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + u_i, \quad \mathbb{E}\left[u_i \middle| z_i\right] = 0, \tag{20}$$

with the outcome variable given by

$$\overline{y}_i := \overline{F}\left(u_i, z_i\right) = \gamma_1\left(x_{i1}, z_i\right) = \gamma_2\left(x_{i2}, z_i\right),$$

replaced by its plug-in estimator

$$\tilde{y}_i := \begin{cases} \hat{\gamma}_1\left(x_{i1}, z_i\right), & \text{for } d_i = 1, \\ \hat{\gamma}_2\left(x_{i2}, z_i\right), & \text{for } d_i = 2, \end{cases}$$

Again using z_i as IVs, estimate α by

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

The difference between $\hat{\alpha}$ and $\hat{\alpha}^*$ lies in the outcome variable being used for the IV regression: $\hat{\alpha}$ is based on the raw output y_i, while $\hat{\alpha}^*$ is based on the estimated conditionally expected output $\overline{y}_i$. As we will show below, $\hat{\alpha}^*$ is in fact asymptotically more efficient than $\hat{\alpha}$.

Theorem 3 (Asymptotic Normality of $\hat{\alpha}^*$)**.** *Define*

$$g^*\left(w_i, \tilde{\alpha}, \tilde{\gamma}\right) := \begin{cases} \overline{z}_i\left(\tilde{\gamma}_1\left(x_{i1}\right) - \tilde{\alpha}_0 - \tilde{\alpha}_1 x_{i1} - \tilde{\alpha}_2 \tilde{\gamma}_2^{-1}\left(\tilde{\gamma}_1\left(x_{i1}\right)\right)\right) & \textit{for } d_i = 1, \\ \overline{z}_i\left(\tilde{\gamma}_2\left(x_{i2}\right) - \tilde{\alpha}_0 - \tilde{\alpha}_2 x_{i2} - \tilde{\alpha}_1 \tilde{\gamma}_1^{-1}\left(\tilde{\gamma}_2\left(x_{i2}\right)\right)\right) & \textit{for } d_i = 2, \end{cases}$$

and $g^*\left(w_i, \tilde{\gamma}\right)$ as well as G^*similarly as in Section 3.1.3. Define

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

Under Assumptions 1-10 with G, φ replaced by G^*, φ^* whenever applicable,

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

where $\Sigma^* := \Sigma_{zx}^{-1}\Omega^*\Sigma_{xz}^{-1}$ *and*

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

The proof is very similar to that of Theorem 2, and is presented in Appendix B.5.

Next, we compare the asymptotic variances of $\hat{\alpha}^*$ and $\hat{\alpha}$, and show that $\hat{\alpha}^*$ is in fact asymptotically more efficient.

Theorem 4 ($\hat{\alpha}^*$ is Asymptotically More Efficient than $\hat{\alpha}$)**.** $\Omega - \Omega^*$ *is positive definite, i.e.,* $\hat{\alpha}^*$ *is asymptotically more efficient than* $\hat{\alpha}$*.*

The proof is in Appendix B.6. Here we discuss the intuition of Theorem 4. The error term for the IV regression with the raw outcome y_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 outcome $\overline{y}_i$ is used instead. Even though we do not observe $\overline{y}_i$ and must use an estimator $\tilde{y}_i = \hat{\gamma}_1\left(x_{i1}\right)$ or $\tilde{y}_i = \hat{\gamma}_2\left(x_{i2}\right)$, 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 "$1 + \varphi\left(w_i\right)$" in (i): the "1" comes from the one "raw" share of error ϵ_i embedded in each y_i that we use as the outcome variable, while "$\varphi\left(w_i\right)$" essentially captures the share of influence of the first-step estimation error $\hat{\gamma} - \gamma$ due to ϵ_i. Together, we have

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

while the corresponding multiplier φ^* on ϵ_i in (ii) is essentially the same except that "$1-\lambda_1\frac{\alpha_2}{\gamma_2^{'}}$" becomes "$\lambda_1-\lambda_1\frac{\alpha_2}{\gamma_2^{'}}$" and "$1-\lambda_2\frac{\alpha_1}{\gamma_1^{'}}$" becomes "$\lambda_2-\lambda_2\frac{\alpha_1}{\gamma_1^{'}}$". Since $\lambda_1, \lambda_2 < 1$, the overall multiplier on ϵ_i becomes smaller in magnitude[27]. Essentially, by using the estimated conditional expected output $\tilde{y}_i$, the raw "1" share of ϵ_i in y_i is moved into the first-stage estimation error of $\overline{y}_i$, 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 4 does not directly relate to the theory of semiparametric efficiency bounds, such as in Ackerberg et al. (2014),

[27]Note that $\alpha_1/\gamma_1^{'} \leq 1$ and $\alpha_2/\gamma_2^{'} \leq 1$ by equation (8).

which is about asymptotic efficiency of semiparametric estimators under a given criterion function. In fact, by Ackerberg et al. (2014), both estimators based on y_i and $\tilde{y}_i$ attain their corresponding semiparametric efficiency bounds with respect to their different criterion functions g and g^*. Theorem 4, however, is a comparison across the two criterion functions g and g^*: it essentially states that the asymptotically efficient estimator under g^* is even more efficient than the efficient estimator under g.

B.5 Proof of Theorem 3

Proof. We adapt the proof of Theorem 2 above with

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

with $\mathbb{E}\left[g^*\left(w_i, \gamma\right)\right] = \mathbb{E}\left[\overline{z}_i\left(\gamma_j\left(x_{ij}\right) - \alpha_0 - \alpha_j x_{ij} - \alpha_k \gamma_k^{-1}\left(\gamma_j\left(x_{ij}\right)\right)\right)\right] = \mathbb{E}\left[\overline{z}_i u_i\right] = \mathbf{0}$ and $\frac{1}{N}\sum_{i=1}^{N} g\left(z, \hat{\alpha}^*, \hat{\gamma}\right) = \mathbf{0}$.

By the chain rule,

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

and

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

By Proposition 4 of Newey (1994), with

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

we have

$$\overline{z}_i \left(\lambda_1 \left(1 - \frac{\alpha_2}{\gamma_2^{'}} \right) + \lambda_2 \frac{\alpha_1}{\gamma_1^{'}},\ \lambda_1 \frac{\alpha_2}{\gamma_2^{'}} + \lambda_2 \left(1 - \frac{\alpha_1}{\gamma_1^{'}} \right) \right) \left(\begin{array}{c} d_{i1} \left(y_i - \gamma_1 \left(x_{i1} \right) \right) \\ d_{i2} \left(y_i - \gamma_2 \left(x_{i2} \right) \right) \end{array} \right) \equiv \varphi^* \left(w_i \right) \overline{z}_i \epsilon_i,$$

and by Assumption 10

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

Hence, we have

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

where

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

giving

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

□

B.6 Proof of Theorem 4

Proof. By (7), we have

$$\frac{\partial}{\partial c} \gamma_j \left(c; z \right) = \alpha_j + \alpha_k x_k^{'} \frac{1}{x_j^{'}} + \frac{1}{x_j^{'}} > \alpha_j,$$

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,

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

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

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

is positive definite. □

C Robustness Check for First Application

Although most pharmacies in our sample have one manager and one pharmacist, there are a few pharmacies with more than one employee pharmacist. For this subset of pharmacies, we compute the total hours worked by employee pharmacists by multiplying the reported hours worked from an employee by the number of employees. Then, the second imputation step is applied based on the total hours worked by all employees. In this process, we implicitly assume the labor hours from two different employees are perfect substitutes. As a robustness check, we also estimate a version of production function which has an elasticity of substitution between the hours worked by different employees equal to one. Table 12 summarizes this version of the estimation result. The estimated parameters show that employees become slightly less productive at both independents and chains compared to our baseline estimation, but in general our estimation result is robust to how we treat employee inputs from pharmacies with more than one employee.

Table 12: Using $N_2 * log(x_2)$ instead of $log(N_2 * H_2)$

	Independent		Chain	
	Observed Outputs	Expected Outputs	Observed Outputs	Expected Outputs
α_0	5.493	5.888	3.409	4.201
	(0.527)	(0.270)	(1.656)	(0.972)
α_1	0.258	0.178	0.878	0.719
	(0.121)	(0.057)	(0.446)	(0.261)
α_2	0.033	0.017	0.092	0.056
	(0.021)	(0.014)	(0.039)	(0.022)
Nobs	144	144	188	188
First-stage F for x_1	10.066	10.066	10.199	10.199
First-stage F for x_2	12.360	12.360	3.210	3.210

D Inter Vivos Gifts

Consider an example with a married couple and two parental households, $j = 1, 2$, whose wealth levels are respectively m_1 and m_2, which is based on Bergstrom, Blume, and Varian (1986). Parents are altruistic toward their married offspring but not toward that offspring's spouse. Parental household j has utility

$$u_j(g_j) = \ln(m_j - g_j) + \mu \ln(g_1 + g_2)$$

where g_j is the married couple's gift from parental household j and μ is the probability that both parental households think the children's marriage will endure. This leads to a noncooperative game between the two parental households since the incentive for either household to gift the offspring couple diminishes as the other parental household gives more. This is a game of strategic substitutes. The Nash equilibrium of this game between the two parental households is

$$g_1^* = \frac{(1+\mu)\, m_1 - m_2}{2+\mu}, \quad g_2^* = \frac{(1+\mu)\, m_2 - m_1}{2+\mu}.$$

There is a unique Nash equilibrium for any μ for any wealth levels for the two households that are not "too" different. Both g_1 and g_2 are strictly increasing in the shock μ, and hence the outcome is strictly increasing in μ. Finally, we can interpret the length of time the marriage survives as a measure of the durability of the marriage.