Production Function Estimation Using Cross Sectional Data: A Partial Identification Approach

Tadashi Sonoda
Nagoya University
sonoda@soec.nagoya-u.ac.jp

Ashok Mishra
Louisiana State University
AMishra@agcenter.lsu.edu

Selected Paper prepared for presentation at the 2015 Agricultural and Applied Economics Association and Western Agricultural Economics Association Annual Meeting, San Francisco, CA, July 26-28

Since the seminal work of Marschak and Andrews (1994), studies on production function estimation have developed by constant efforts to reexamine how to deal with endogenous labor input. Because of unobserved productivity or managerial ability of producers, these studies gave up estimating production functions using cross sectional data and instead proposed estimation methods based on panel data (e.g., Griliches and Mairesse, 1998). Popular methods include fixed effects (FE) estimation (e.g., Mundlak, 1961), generalized method of moments (GMM) using lags or differences of inputs and outputs as instrumental variables (e.g., Blundell and Bond, 1999), and proxy variable approaches of Olley and Pakes (1996) and Levinsohn and Petrin (2003).

Although panel data might be available for more regions than before, there are still many other regions for which only cross sectional data are available for production function estimation. Furthermore, those cross sectional data might contain much more observations and information than typical panel data can provide. In this case, we hope to find an appropriate way to estimate a production function using cross sectional data.

Moreover, those popular estimation methods for panel data tend to impose strong and untestable assumptions. FE estimation assumes unobserved productivity to be constant over time and the estimated production elasticity of capital tends to be biased downward. GMM estimation assumes that instrumental variables (IVs) have sufficient correlation with endogenous regressors and no correlation with the error term, but full investigation of the latter is impossible. Olley-Pakes and Levinsohn-Petrin methods convert unobservable productivity into observable proxies (capital investment or intermediate inputs) and require that the proxies take only positive values, increase strictly with productivity, and should not include measurement errors.

These observations show that it is important to propose a practical method for production function estimation using cross sectional data and weaker assumptions. For this purpose, we adopt and improve a partial identification approach of Nevo and Rosen (2012) to

estimate the upper and lower bounds of production elasticities by the imperfect instrumental variables (IIV) method. The most attractive feature of IIV method is to allow correlation between labor inputs and the error term, which is one of the most difficult issues and paves the way to production function estimation using cross sectional data. Although the IIV method enables us to estimate only intervals of production elasticities, we believe that their intervals can be more helpful than their biased point estimates, as emphasized by other studies taking a partial identification approach.

The second section explains data used in our empirical analysis. The third section estimates Cobb-Douglas production functions using the methods of ordinary least squares (OLS) and instrumental variables (IVs) for comparison purpose. The fourth section introduces the IIV method, explains how to choose IIVs, and proposes an alternative way to determine the weight of combining two IIVs. It then shows results of the estimated upper and lower bounds of the production elasticity of labor. The final section concludes the paper.

Data

Our empirical analysis uses data of rural households in the Chinese Household Income Project (CHIP) survey in 2002 (Li, 2002), which covers 22 out of 31 provinces in China (see Gustafsson, Li, and Sicular (2008) and Knight, Deng, and Li (2011) for a detailed description of the survey). We specifically examine 4174 typical households for which data on relevant variables are not missing and the following conditions are satisfied: 1) value-added, costs of producing crops or livestock, and cultivated areas are all positive, 2) the household head is a married male, 3) both the household head and his wife work on their own farm, and 4) sample villages include at least two households. We also use village-level data from the Administrative Village Questionnaire annexed to the CHIP survey.

Output (*value_added*) is defined as the difference between value of output and costs of producing grain, economic crops, and livestock products. Labor (*farmlabor*) is measured by total farm work hours of family members to produce crops (grain and economic crops) and livestock products. Farm capital (*capital*) is total value of large and medium sized farm tools, farm machinery and equipment, livestock, transportation machinery and equipment, and structures used for production. Land (*land*) is measured by cultivated areas of own and rented land.

Table 1 presents sample means of the output and inputs. To allow for regional differences in farm production and apply our IIV method to different data, we classify the 22 provinces into the eastern, central, and western regions. Mean of value-added is 4637, 4800, and 4054 yuan (1 yuan = 0.16 dollar) for the eastern, central, and western regions, respectively. In the eastern region, where economy has been developing most rapidly, farm work hours are the shortest (2351), cultivated land is the smallest (6.4 mu = 0.4 ha), and farm capital is the largest (4639 yuan) of the three regions. These results reflect higher incomes and more participation in wage work for farm households in this region. In the western region, where economy has been developing most slowly, farm work hours are 37% longer and farm capital is 22% smaller than those in the eastern region. In addition to fewer opportunities for wage work, farm households in this region produce more livestock products, which partly explains their longer hours on farm and smaller cultivated land. In the central region, value share of grain in total farm products is 52% and cultivated land is the largest (8.8 mu = 0.6 ha). Although farm mechanization progressed and it helped shorten farm work hours, farm capital in this region is similar to that in the western region and it is 20% smaller than the one in the eastern region.

Production Function Estimation Using OLS and IV Methods

We specify and estimate a Cobb-Douglas production function:

$$ln(value\_added) = \alpha_0 + \alpha_1 ln(farmlabor) + \alpha_2 ln(capital) + \alpha_3 ln(land) + \alpha_4 irr\_share + \alpha_5 educ\_head + \alpha_6 age\_head + \alpha_7 age\_head^2 + u, \quad (1)$$

where u denotes an error term. Following Jacoby (1993), production function (1) includes four variables to control for household fixed effects to some extent. Schooling years of the household head ($educ\_head$), his age (age_head), and its square are used to control for managerial ability. The share of irrigated land (irr_share) is used to control for land quality. Sample means of these variables are shown in Table 1. We always include province dummy variables when estimating the production function (1).

We first use OLS to estimate equation (1) for the three regions. Table 2 presents the results. Production elasticities of labor, capital, and land are estimated statistically significant and positive for all regions. Production elasticity of capital is lower than 0.1 and that of land is nearly 0.45 for all regions. Production elasticity of labor varies greatly with regions: it is 0.43, 0.20, and 0.16 for the eastern, central, and western regions, respectively. The share of irrigated land and age of the household head have statistically significant effects for some cases, while education of the household head does not have significant effects.

Our result is most appropriately compared with Yang (1997a) because he uses OLS to estimate a Cobb-Douglas production function which has value added for crops and livestock as outputs and includes the household head's education as a shift factor. Using cross sectional data on 197 households in Sichuan, a province in western China, he estimates production elasticities of labor, capital, and land at 0.23, 0.04, and 0.49, respectively. These estimates are similar to ours for the western region, 0.16, 0.08, and 0.44.

We next estimate equation (1) using the IV method. We assume that only $farmlabor$ is potentially endogenous throughout this study, which seems plausible because we use cross sectional data and only 8% of households rent land from other households in our sample.

We simply choose two IV sets similar to Yang (1997b) and Jacoby (1993) because candidates for IVs are limited for the cross sectional data. Set 1 includes only the number of adults in the household. Set 2 includes ten variables: a dummy variable for drinking water from tap (*tapwater*), a dummy variable for using firewood for fuel (*firewood*), share of own cultivated land (*ownland*), number of adult males and females (*num_adult_male, num_adult_female*), number of children younger than 6 years old (*num_childyt6*), number of children between 6 and 15 years old (*num_childge6*), a village dummy variable of large population (*vill_large*), village-level daily wage (*vill_wage*), and village-level rice price (*vill_p_rice*). Sample means of the variables are shown in Table 1.

Table 3 presents the results. IV estimation with set 2 yields similar production elasticities to those obtained by OLS. In particular, production elasticity of labor is estimated at 0.49, 0.19, and 0.18 for the eastern, central, and western regions, respectively. IV estimation with set 1 yields slightly different production elasticities from those obtained by OLS particularly for the western region: production elasticity of labor is estimated at 0.35 for this region.

Table 3 also presents an F statistic to test significance of instruments (F stat.) and the adjusted coefficient of determination ($\bar{R}^2$) in the first stage regression, and a χ^2 statistic to test overidentifying restrictions (OIR stat.). The F statistic for IVs in set 2 is smaller than the critical value of Stock and Yogo (2005), which indicates weakness of these IVs. Furthermore, the OIR statistic for these IVs is greater than the critical value of χ^2 statistic for all regions, which indicates rejection of their orthogonality. On the other hand, the number of adults (the only IV in set 1) is not weak, although we cannot test its orthogonality. Finally, Table 3 presents Durbin-Wu-Hausman statistic (D-W-H stat.) to test exogeneity of $farmlabor$ in equation (1), which is not rejected at the 5% level for all cases.

Consequently, we use the OLS estimates as a benchmark in the analysis below. However, it should be noted that the D-W-H test critically depends on untestable assumptions: at

least one instrument is indeed exogenous for each of the two IV sets. We next ask validity of these assumptions using IIV methods.

Production Function Estimation Using IIV Methods

Basic Theoretical Conditions for IIVs

To focus on endogeneity of labor, we rewrite production function (1) as

$$Y = \alpha X + \boldsymbol{W}'\boldsymbol{\delta} + u, \tag{2}$$

where Y, X, $\boldsymbol{W}$, and u respectively denote $ln(value\_added)$, $ln(farmlabor)$, the vector of exogenous variables, and the error term. As in Nevo and Rosen (2012), our main concern is identification and estimation of parameter α, production elasticity of labor, which can be safely assumed to be non-negative. To further focus on α, we rewrite equation (2) as a simple regression model:

$$\tilde{Y} = \alpha\tilde{X} + u, \quad \alpha \geq 0 \tag{3}$$

Variable $\tilde{X}$ $(\tilde{Y})$ represents residuals in the population regression of X (Y) on $\boldsymbol{W}$ and therefore $E\tilde{X} = E\tilde{Y} = 0$.

According to the IIV method, the upper and lower bounds of α are given by OLS estimator, α_{OLS}, or IV estimator, $\alpha_{IV}(Z)$, applied to equation (3).

$$\alpha_{OLS} = \sigma_{\tilde{x}\tilde{y}}/\sigma_{\tilde{x}\tilde{x}}, \quad \alpha_{IV}(Z) = \sigma_{z\tilde{y}}/\sigma_{z\tilde{x}} \tag{4}$$

where $\sigma_{ab} = cov(A, B)$ for random variables A and B. For simplicity, we refer to α_{OLS} and $\alpha_{IV}(Z)$ as OLS and IV estimators of α, respectively, rather than their probability limits.

Variable Z is called an IIV if it satisfies the following conditions:

(C1) $\rho_{z\tilde{x}} \neq 0$

(C2) $\rho_{xu}\rho_{zu} \geq 0$

where ρ_{ab} denotes correlation coefficient between random variables A and B. Condition

(C1) requires Z to be correlated with endogenous regressor $\tilde{X}$. Condition (C2) requires that correlation between Z and the error term u cannot have the opposite sign to correlation between X and u.

For production function estimation using cross sectional data, these conditions can be made more specific. First, we assume positive correlation between X and u, which has been supported by most studies of production function estimation (e.g., Olley and Pakes,1996). It is justified if favorable productivity shocks cause a neutrally upward shift of the production function and if this shift expands the demand for labor. In this case, (C2) is rewritten as

(C2') $\rho_{xu} > 0$ and $\rho_{zu} \geq 0$

Second, we assume availability of IIVs which have positive partial correlation with farm work hours.

(C1') $\rho_{z\tilde{x}} > 0$

This condition is easily checked because X ($\tilde{X}$) and Z are observable (or estimable). To exemplify a potential variable satisfying both (C1') and (C2'), we can invoke the demand for intermediate inputs, a proxy in the Levinsohn-Petrin method. This demand is likely to respond positively to productivity shocks for a similar reason for the positive correlation between X and u. Furthermore, X and Z are likely to move in the same direction in response to productivity shocks, given fixed amount of land and/or capital, which is likely to support positive correlation between $\tilde{X}$ and Z.

In addition to (C1') and (C2'), a narrower interval of α might be obtained if correlation between Z and u is weaker than correlation between X and u.

(C3) $|\rho_{xu}| \geq |\rho_{zu}|$

We will examine plausibility of (C3) for each IIV candidate to be introduced below.

Conditions (C1'), (C2'), and (C3) for IIVs are weaker than conditions required by other estimation methods if we appropriately choose Z to satisfy relations in these conditions.

The IV or GMM method requires the orthogonality condition, $\rho_{zu} = 0$. The Olley-Pakes or Levinsohn-Petrin method requires that the proxy for productivity (capital investment or intermediate inputs) takes only positive values, increases strictly with productivity, and should not include measurement errors. The cost of this advantage of IIV method is unavailability of a point estimate of α. Given conditions (C1') and (C2'), we find only the upper bound of α using a single IIV, and we find the lower and upper bounds of α using two IIVs together. To minimize this cost, we try to find a narrower range of α by choosing better IIVs and by deriving useful relations to impose plausible assumptions.

Choice of IIVs and Additional Requirements for Them

Following Nevo and Rosen (2012), we will use two IIVs together to derive two-sided bounds of α under conditions (C1'), (C2') and (C3). A practical issue to apply their method is how to find a better pair of IIVs among candidates satisfying these conditions. We begin with two candidates which seem to satisfy the three conditions and were used as a potentially invalid IV or a proxy variable for productivity shocks.

The first candidate is the number of household members who work on their own farm. It has positive partial correlation with farm work hours ($\rho_{z\tilde{x}} > 0$) as shown below. The number of farm workers might not respond to productivity shocks ($\rho_{zu} = 0$). However, it can respond positively to favorable productivity shocks ($\rho_{zu} > 0$) because the household head may ask help of his/her family members who usually do not work on farm (his/her parents or children). Furthermore, we expect the number of farm workers to respond more weakly to productivity shocks than farm work hours, as condition (C3) requires.

The second candidate is cost of intermediate inputs in farm production. As explained above, it is likely to have positive partial correlation with farm work hours ($\rho_{z\tilde{x}} > 0$) and respond positively to favorable productivity shocks ($\rho_{zu} > 0$). Furthermore, farm

households (particularly in developing countries) might adjust their demand for labor more flexibly than their demand for intermediate inputs in response to productivity shocks because the latter often needs immediate cash payment. We will return to plausibility of condition (C3) for this candidate later.

Allowing for data availability, our IIV candidates include the number of male and female adults who work on their own farm (num_farmer_male and num_farmer_female), the sum of these adults (num_farmer), and the logarithm of total costs of producing crops and raising livestock ($ln(pcost\_all)$). Sample means of these variables are presented in Table 1. Although these candidates might satisfy the three theoretical conditions, they might not be good enough to obtain a narrow interval of α.

To explain this point, we review choice of IIVs by Nevo and Rosen (2012). They estimate a market share function for various cereal brands, which has the price $p_{j,t}$ of cereal brand j in market t as an endogenous regressor ($t = B$ for Boston and $t = S$ for San Francisco). Similarly to our case, their IIVs satisfy conditions (C1') and (C2') above and they use two IIVs together to derive two-sided bounds of parameters of the share function. One is the price $p_{j,r}$ of cereal brand j in the other market $(r = B, S;\ r \neq t)$. The other is its average price $\bar{p}_{j,R}$ in other cities $k\ (\neq t)$ of the same region R, where R is New England for $t = B$ and northern California for $t = S$. Focusing on partial correlation between these IIVs and the endogenous price, they report it to be 0.81 for $p_{j,r}$ and 0.48 for and $\bar{p}_{j,R}$.

Their choice of IIVs suggests additional requirements for IIVs. First, correlation of each IIV with $\tilde{X}$ must be not only statistically significant but also moderately high. Second, for a pair of IIVs, one must have relatively high correlation with $\tilde{X}$ and the other must have relatively low correlation with $\tilde{X}$. Columns "hh vars" (household-level variables) in Table 4 present correlation $\rho_{z\tilde{x}}$ between $\tilde{X}$ and Z for the three regions, where t-value for Z in the regression of $\tilde{X}$ on Z is shown in parentheses. The table shows that $\rho_{z\tilde{x}}$ is statistically significant and ranges from 0.12 to 0.24 for the four candidates in all regions. Consequently,

these candidates barely satisfy the first requirement, whereas few pairs of them seem to satisfy the second requirement.

To introduce another candidate which satisfies the three conditions for IIVs and has higher correlation with $\tilde{X}$, we recall relation between the prices $p_{j,t}$ and $\bar{p}_{j,R}$ in the above application. Let X_j denote the logarithm of farm work hours for household j and let h $(\neq j)$ index other households in the village where household j lives, with N_j denoting the number of these households. By defining the average of X_h for other households in the village, $\bar{X}_{-j} \equiv (N_j - 1)^{-1} \sum_{h \neq j} X_h$, it proves to satisfy the conditions (C1'), (C2'), and (C3) under plausible assumptions. Furthermore, since the logarithm of production cost, $W_j \equiv ln(pcost\_all_j)$, might not satisfy condition (C3), we replace it with the village average $\bar{W}_{-j}$, which is constructed in a similar way to $\bar{X}_{-j}$.

Columns "v. mean" (village mean) in Table 4 present correlation between $\tilde{X}$ and variables averaged over other households in the village. As expected, $\bar{X}_{-j}$ in the row $ln(farmlabor)$ has high correlation with $\tilde{X}$: it is 0.57, 0.59, and 0.50 for the eastern, central, and western regions. This high correlation will help our IIV candidates satisfy the second requirement. Moreover, correlation coefficient of $\bar{W}_{-j}$ with $\tilde{X}$, which is shown in the row $ln(pcost\_all)$, is 0.17, 0.11, and 0.05 for the three regions and seems to satisfy the first requirement at least for the eastern and central regions. In summary, we use num_farmer, num_farmer_male, num_farmer_female, $\bar{W}_{-j}$, and $\bar{X}_{-j}$ for our IIVs in the subsequent analysis.

Identification and Estimation with a Single IIV

To identify and estimate production elasticity of labor, α, we first use a single IIV and apply Proposition 2 of Nevo and Rosen (2012). In addition to (C1'), (C2'), and (C3), we assume a condition which can be checked by the data:

(C4) $\tau_{zx} \equiv (\sigma_z \sigma_{x\tilde{x}} - \sigma_x \sigma_{z\tilde{x}})\sigma_{z\tilde{x}} > 0$

Under the four conditions, only the upper bound of α is identified as

$$\alpha \leq \overline{\alpha} = min\{\alpha_{IV}(Z), \alpha_{IV}(V)\}, \tag{5}$$

where $\alpha_{IV}(V) = \sigma_{v\tilde{y}}/\sigma_{v\tilde{x}}$, $V = \sigma_x Z - \sigma_z X$, and σ_a denotes standard deviation of random variable A. Using definition of V, $\alpha_{IV}(V)$ is rewritten as

$$\alpha_{IV}(V) = \beta\, \alpha_{IV}(Z) + (1-\beta)\alpha_{OLS} = \alpha_{OLS} + \beta\{\alpha_{IV}(Z) - \alpha_{OLS}\}, \tag{6}$$

where $\beta \equiv -\sigma_x \sigma_{z\tilde{x}}^2/\tau_{zx}$. Since condition (C4) means $\beta < 0$, relation (6) implies that

$$\begin{aligned} \overline{\alpha} &= \alpha_{IV}(V) \quad \text{if} \quad \alpha_{OLS} \leq \alpha_{IV}(Z), \\ &= \alpha_{IV}(Z) \quad \text{if} \quad \alpha_{OLS} > \alpha_{IV}(Z). \end{aligned} \tag{7}$$

A seemingly natural estimator of $\overline{\alpha}$ might be $min\{\hat{\alpha}_{IV}(Z), \hat{\alpha}_{IV}(V)\}$, where $\hat{\alpha}_{IV}(Z)$ and $\hat{\alpha}_{IV}(V)$ are obtained by replacing σ_a and σ_{ab} by their sample analogues. However, discussion of Chernozhukov, Lee, and Rosen (2013) shows that the estimator $min\{\hat{\alpha}_{IV}(Z), \hat{\alpha}_{IV}(V)\}$ tends to be downward biased in finite samples and does not account for unequal sampling errors between $\hat{\alpha}_{IV}(Z)$ and $\hat{\alpha}_{IV}(V)$. Following Chernozhukov et al., we use a simulation-based estimator of $\overline{\alpha}$:

$$\hat{\overline{\alpha}}(p) = min_{q \in Q}\{\hat{\alpha}_{IV}(q) + k_{\hat{Q}}(p)s(q)\}, \quad q \in Q = \{Z, V\} \tag{8}$$

In particular, $\hat{\overline{\alpha}}(0.50)$ represents a half-median unbiased estimator of $\overline{\alpha}$, which is comparable with a point estimate of $\hat{\alpha}_{OLS}$. Furthermore, $\alpha \leq \hat{\overline{\alpha}}(0.975)$ provides a 97.5% confidence interval of the identified region (5), which is comparable with a confidence interval $\alpha \leq \hat{\alpha}_{OLS} + c_{0.975}\, se(\hat{\alpha}_{OLS})$ for OLS estimator ($c_{0.975}$: 97.5th percentile in the standard normal distribution, $se(\hat{\alpha}_{OLS})$: standard error of $\hat{\alpha}_{OLS}$). In equation (8), principal critical value, $k_{\hat{Q}}(p)$, represents the pth quantile of the distribution of $max_{q \in \hat{Q}}\{Z^*(q)\}$, where $Z^*(q)$ is the weighted average of two standard normal variables. The set $\hat{Q}$ is an adaptive inequality selector developed by Chernozhukov et al. and $s(q)$ denotes an estimator of the normalizing factor for $Z^*(q)$ (see Appendix A for detailed computation of (8)).

Table 5 presents estimates of $\hat{\alpha}_{OLS}$, $\hat{\alpha}_{IV}(Z)$, $\hat{\alpha}_{IV}(V)$, and $\widehat{\tilde{\alpha}}(0.50)$, where standard errors are shown in parentheses and estimates of $\widehat{\tilde{\alpha}}(0.975)$ are shown in brackets. The five IIVs are $C_1 = num\_farmer$, $C_2 = num\_farmer\_male$, $C_3 = num\_farmer\_female$, $C_4 = \bar{W}_{-j}$ and $C_5 = \bar{X}_{-j}$. The table also presents the weight β in relation (6), which is negative for all cases and verifies the condition (C4).

We can interpret estimates of $\widehat{\tilde{\alpha}}(0.50)$ in two steps. First, we examine $\tilde{\alpha} = min\{\hat{\alpha}_{IV}(Z), \hat{\alpha}_{IV}(V)\}$ by ignoring sampling errors. In the central and western regions, $\tilde{\alpha} = \hat{\alpha}_{IV}(V)$ for $Z = C_1, \cdots, C_4$ and $\tilde{\alpha} = \hat{\alpha}_{IV}(Z)$ for $Z = C_5$: $\tilde{\alpha}$ is estimated between 0.05 and 0.19 for the central region and between 0.04 and 0.13 for the western region, both of which are slightly lower than the corresponding OLS estimates. In the eastern region, $\tilde{\alpha} = \hat{\alpha}_{IV}(Z)$ for $Z = C_1, \cdots, C_3$ and $\tilde{\alpha} = \hat{\alpha}_{IV}(V)$ for the other IIVs: $\tilde{\alpha}$ is estimated between 0.09 and 0.35, which is much lower than the OLS estimate. Next, if we take account of sampling errors, difference between $\widehat{\tilde{\alpha}}(0.50)$ and $\tilde{\alpha}$ is found to be positive for all cases but it is small for most cases: the difference is nearly 0.10 for $Z = C_1, \cdots, C_3$ in the eastern region but it is at most 0.04 for the other cases. Consequently, $\widehat{\tilde{\alpha}}(0.50)$ is lower than the corresponding OLS estimate for most cases, in which cases $\widehat{\tilde{\alpha}}(0.50)$ is estimated between 0.22 and 0.33 in the eastern region, between 0.07 and 0.19 in the central region, and between 0.09 and 0.14 in the western region.

Now, we examine estimates of $\widehat{\tilde{\alpha}}(0.975)$. The difference between $\widehat{\tilde{\alpha}}(0.975)$ and $\widehat{\tilde{\alpha}}(0.50)$ is composed of the difference between the two critical values, $k_{\hat{Q}}(0.975)$ and $k_{\hat{Q}}(0.50)$, and the "standard error" $s(q)$ of chosen IV estimator $\hat{\alpha}_{IV}(q)$. The difference between the two critical values is approximately 1.80 for all IIVs in all regions. The "standard error" $s(Z)$ varies with IIVs and regions, but $s(V)$ is smaller than 0.06 for all IIVs in all regions. These relations imply that the difference between $\widehat{\tilde{\alpha}}(0.975)$ and $\widehat{\tilde{\alpha}}(0.50)$ is at most 0.10 if $\hat{\alpha}_{IV}(V) + k_{\hat{Q}}(0.975)s(V)$ is smaller than $\hat{\alpha}_{IV}(Z) + k_{\hat{Q}}(0.975)s(Z)$ and it exceeds 0.10 otherwise.

In the central and western regions, $\hat{\bar{\alpha}}(0.975)$ is greater than $\hat{\bar{\alpha}}(0.50)$ at most by 0.10 for all cases but one: $\hat{\bar{\alpha}}(0.975)$ is between 0.13 and 0.31 in the central region and it is between 0.21 and 0.26 in the western region. On the other hand, 97.5% confidence intervals computed from the OLS estimates are $\alpha \leq 0.25$ in the central region and $\alpha \leq 0.23$ in the western region. Consequently, $\alpha \leq \hat{\alpha}_{OLS} + c_{0.975}\, se(\hat{\alpha}_{OLS})$ cannot be partly contained in $\alpha \leq \hat{\bar{\alpha}}(0.975)$ for $Z = C_1, C_3, C_4$ in the central region and for $Z = C_3, C_4$ in the western region.

In the eastern region, $\hat{\bar{\alpha}}(0.975)$ is estimated between 0.53 and 0.57 for $Z = C_1, \cdots, C_3$ because $\hat{\alpha}_{IV}(Z) + k_{\hat{Q}}(0.975)s(Z)$ is smaller or because $\hat{\alpha}_{IV}(V)$ is much greater than $\hat{\alpha}_{IV}(Z)$. Nonetheless, $\hat{\bar{\alpha}}(0.975)$ exceeds $\hat{\bar{\alpha}}(0.50)$ at most by 0.10 and it is estimated at 0.40 for $Z = C_4$ and $Z = C_5$. On the other hand, 97.5% confidence intervals computed from the OLS estimates are $\alpha \leq 0.48$. Consequently, $\alpha \leq \hat{\alpha}_{OLS} + c_{0.975}\, se(\hat{\alpha}_{OLS})$ cannot be partly contained in $\alpha \leq \hat{\bar{\alpha}}(0.975)$ for $Z = C_4, C_5$.

Identification and Estimation with Paired IIVs

To obtain a two-sided bound of α using a pair (Z_1, Z_2) of IIVs, we assume that each Z_j $(j = 1, 2)$ satisfies (C1'), (C2'), (C3), and (C4). We also assume a condition which can be checked by the data:

(C5) $\tau_{xyz} \equiv \sigma_{z_1\tilde{y}}\sigma_{z_2\tilde{x}} - \sigma_{z_2\tilde{y}}\sigma_{z_1\tilde{x}} < 0$

Combined with (C1') for $Z = Z_j$, condition (C5) implies that

$$\alpha_{IV}(Z_1) < \alpha_{IV}(Z_2) \tag{9}$$

This is not a strong requirement because we can choose Z_1 and Z_2 to satisfy it. Furthermore, in place of (C2'), each Z_j is assumed positively correlated with u.

(C2") $\rho_{xu} > 0$ and $\rho_{zu} > 0$

This assumption is not strong because our difficulty in production function estimation

using cross sectional data is unavailability of valid instruments.

Given these conditions and applying Proposition 5 and Lemma 2 of Nevo and Rosen (2012), we find the following two-sided bounds of α for some values of γ which satisfy $0 < \gamma < 1$, $\sigma_{\omega(\gamma)\tilde{x}} < 0$, and $\sigma_{\omega(\gamma)u} \geq 0$:

$$\underline{\alpha} \leq \alpha \leq \overline{\alpha} \tag{10}$$

$$\underline{\alpha} = \alpha_{IV}(\omega(\gamma))$$

$$\overline{\alpha} = min\{\alpha_{IV}(Z_1), \alpha_{IV}(V_1), \alpha_{IV}(V_2), \alpha_{IV}(V^*(\gamma))\},$$

where $\omega(\gamma) = \gamma Z_2 - (1-\gamma)Z_1$, $V_j = \sigma_x Z_j - \sigma_{z_j} X$, and $V^*(\gamma) = \sigma_x \omega(\gamma) - \sigma_{\omega(\gamma)} X$. We use relation (9) to exclude $\alpha_{IV}(Z_2)$ from candidates of the upper bound $\overline{\alpha}$.

In addition to choosing a better pair of IIVs, another practical issue to estimate the bounds (10) is how to set the value of γ. Nevo and Rosen (2012) set $\gamma = 0.5$ or $\gamma = \gamma_{NR} \equiv \sigma_{z_1}/(\sigma_{z_1} + \sigma_{z_2})$. To determine γ by using more information from data and economic theory, we investigate how the upper and lower bounds of α are related to values of γ. Given the five conditions above, we will show that the choice of $\gamma = 0.5$ or $\gamma = \gamma_{NR}$ can easily cause a negative estimate of the lower bound $\underline{\alpha}$ and that the upper bound $\overline{\alpha}$ might not depend on γ. Furthermore, we will use these results to determine a more plausible value of γ. For this analysis, Figure 1 helps understand relations between estimators of α and values of γ, where the circles in white indicate candidates for $\overline{\alpha}$, the circle in black indicates a candidate for $\underline{\alpha}$, and the circle in gray indicates the OLS estimator.

We first examine the lower bound $\underline{\alpha}$ in relation to values of γ. This bound, $\underline{\alpha} = \alpha_{IV}(\omega(\gamma))$, can be written as follows:

$$\alpha_{IV}(\omega(\gamma)) = \alpha_* + A/(\gamma - \gamma_*), \tag{11}$$

$$\gamma_* = \sigma_{z_1\tilde{x}}/(\sigma_{z_1\tilde{x}} + \sigma_{z_2\tilde{x}}),$$

$$\alpha_* = \gamma_* \alpha_{IV}(Z_1) + (1-\gamma_*)\alpha_{IV}(Z_2),$$

where $A \equiv -\tau_{xyz}/(\sigma_{z_1\tilde{x}} + \sigma_{z_2\tilde{x}})^2$ and it is positive from (C5). Relation (11) shows that the lower bound $\underline{\alpha}$ represents a rectangular hyperbola on the (γ, α) plane, with its asymptotes

given by $\gamma = \gamma_* \in (0,1)$ and $\alpha = \alpha_* \in (\alpha_{IV}(Z_1),\ \alpha_{IV}(Z_2))$. Figure 1 illustrates a typical locus of $\alpha_{IV}(\omega(\gamma))$: it crosses the horizontal axis at $\gamma = \gamma_0$, where $\gamma_0 \equiv \sigma_{z_1\tilde{y}}/(\sigma_{z_1\tilde{y}} + \sigma_{z_2\tilde{y}})$ and it ranges from 0 to γ_*. This figure shows that $\underline{\alpha}$ can take a negative value or an extremely large positive value if γ is set at values close to γ_*. It should be noted that γ_* can be close to 0.5 or γ_{NR} because $\gamma_* = 0.5$ if $\sigma_{\tilde{x}z_1} = \sigma_{\tilde{x}z_2}$ and $\gamma_* = \gamma_{NR}$ if $\sigma_{z_1}:\sigma_{z_2} = \sigma_{\tilde{x}z_1}:\sigma_{\tilde{x}z_2}$.

We next examine the upper bound $\overline{\alpha}$ in relation to values of γ. For this purpose, comparison of the four candidates of $\overline{\alpha}$ is useful. To show a typical case in which we can determine their ranking of the candidates, we assume relations (12) and (13):

$$\alpha_{IV}(Z_j) > \alpha_{OLS} \tag{12}$$

$$\{\alpha_{IV}(Z_1) - \alpha_{OLS}\}/\{\alpha_{IV}(Z_2) - \alpha_{OLS}\} < \beta_2/\beta_1 \text{ if } \alpha_{IV}(Z_1) < \alpha_{IV}(Z_2), \tag{13}$$

where $\tau_{z_jx} \equiv (\sigma_{z_j}\sigma_{x\tilde{x}} - \sigma_x\sigma_{z_j\tilde{x}})\sigma_{z_j\tilde{x}}$ and $\beta_j \equiv -\sigma_x\sigma^2_{z_j\tilde{x}}/\tau_{z_jx}$ are the scalar τ_{zx} and the weight β that are defined for $Z = Z_j$. Relations (7) and (12) imply $\alpha_{IV}(Z_j) > \alpha_{IV}(V_j)$. Relations (6) and (13) imply $\alpha_{IV}(V_1) > \alpha_{IV}(V_2)$. These results mean that $\overline{\alpha} = min\{\alpha_{IV}(V_2),\ \alpha_{IV}(V^*(\gamma))\}$.

To compare the two remaining candidates, we can show that $\alpha_{IV}(V^*(\gamma))$ has the following properties:

1) $\alpha_{IV}(V^*(\gamma))$ is a linear combination of $\alpha_{IV}(\omega(\gamma))$ and α_{OLS}. In particular,
   - 1a) $\alpha_{IV}(V^*(\gamma))$ internally divides $\alpha_{IV}(\omega(\gamma))$ and α_{OLS} for $0 \le \gamma < \gamma_*$ and externally divides them for $\gamma_* < \gamma \le 1$,
   - 1b) $\alpha_{IV}(V^*(0))$ is a weighted average of $\alpha_{IV}(Z_1)$ and α_{OLS},
   - 1c) $\alpha_{IV}(V^*(\gamma_{**})) = \alpha_{IV}(\omega(\gamma_{**})) = \alpha_{OLS}$ with $\gamma_{**} \equiv (\alpha_*\gamma_0 - \alpha_{OLS}\gamma_*)/(\alpha_* - \alpha_{OLS})$,
   - 1d) $\alpha_{IV}(V^*(\gamma_*)) = \alpha_{OLS} - A/B\sigma_{\omega(\gamma_*)}$ with $B \equiv \sigma_{x\tilde{x}}/\sigma_x(\sigma_{z_1\tilde{x}} + \sigma_{z_2\tilde{x}}) > 0$, and
   - 1e) $\alpha_{IV}(V^*(1)) = \alpha_{IV}(V_2)$.
2) $\alpha_{IV}(V^*(\gamma))$ decreases monotonically if $\alpha_* > \alpha_{OLS}$ and if

$$(\partial\sigma_{\omega(\gamma)}/\partial\gamma)(\gamma - \gamma_{**})/\sigma_{\omega(\gamma)} < 1 \quad \text{or} \quad (\sigma^2_{z^+}\gamma_{**} - \sigma_{z^+z_1})\gamma > \sigma_{z^+z_1}\gamma_{**} - \sigma^2_{z_1}, \tag{14}$$

where $Z^+ \equiv Z_1 + Z_2$.

Figure 1 illustrates a typical locus of $\alpha_{IV}(V^*(\gamma))$ which satisfies these properties: it takes a

value between $\alpha_{IV}(Z_1)$ and α_{OLS} at $\gamma = 0$, crosses the lower bound curve $\alpha = \alpha_{IV}(\omega(\gamma))$ at $\gamma = \gamma_{**}$, reaches a point slightly below $\alpha = \alpha_{OLS}$ at $\gamma = \gamma_*$, and finally reaches $\alpha = \alpha_{IV}(V_2)$ at $\gamma = 1$. Consequently, we expect a relation $\alpha_{IV}\big(V^*(\gamma)\big) \geq \alpha_{IV}(V_2)$ and therefore $\overline{\alpha} = \alpha_{IV}(V_2)$.

Using the relations among $\underline{\alpha}$, $\overline{\alpha}$, and γ derived above, we propose an alternative way to set γ which reflects more information from data and economic theory. Lemma 2 of Nevo and Rosen (2012) shows that condition (C5) is equivalent to the following two conditions:

(C5') $\sigma_{\omega(\gamma)\tilde{x}} < 0$ and $\sigma_{\omega(\gamma)u} \geq 0$

Combining (C5') with (C1') and (C2") yields the following interval of γ:

$$\gamma_{***} \leq \gamma < \gamma_*, \quad \gamma_{***} = \sigma_{z_1 u}/(\sigma_{z_1 u} + \sigma_{z_2 u}) \tag{15}$$

Because the lower bound γ_{***} is not estimable, a practical interval of γ is

$$0 \leq \gamma < \gamma_* \tag{16}$$

To obtain a narrower interval of γ, we maintain (16) and add two conditions to exclude unlikely values of γ. First, recalling non-negativity of α in relation (3), we impose the following condition on the lower bound of α:

(C6) $\underline{\alpha} = \alpha_{IV}(\omega(\gamma)) \geq 0$

This condition is solved for γ to derive $\gamma \leq \gamma_0$, where γ_0 is defined above and satisfies $\gamma_{***} \leq \gamma_0 < \gamma_*$. Interval (16) then narrows to

$$0 \leq \gamma < \gamma_0 \tag{17}$$

Second, we assume that the lower bound of α cannot exceed its upper bound.

(C7) $\underline{\alpha} = \alpha_{IV}(\omega(\gamma)) \leq \overline{\alpha}$

We examine two cases to derive the interval of γ from condition (C7). In the case where the upper bound $\overline{\alpha}$ is given by $\alpha_{IV}(Z_1)$ or $\alpha_{IV}(V^*(\gamma))$, we cannot gain any information to restrict γ because (C7) is always satisfied for $0 < \gamma < \gamma_*$. In the case where $\overline{\alpha}$ is given by $\alpha_{IV}(V_1)$ or $\alpha_{IV}(V_2)$, we solve (C7) for γ to find

$$\gamma \geq \gamma_{min} \equiv \Gamma_1/(\Gamma_1 + \Gamma_2) \geq 0, \quad \Gamma_j = \sigma_{z_j \tilde{x}}\{\alpha_{IV}(Z_j) - \overline{\alpha}\} \quad (j = 1, 2) \tag{18}$$

where $\Gamma_j > 0$ under condition (C1').

Using the values of γ in relations (16), (17), and (18), which reflect information of data and economic theory, we set γ in the following way:

$$\begin{aligned} \hat{\gamma} &= \gamma_v \quad \text{if } \overline{\alpha} = \alpha_{IV}(V^*(\gamma_v)) \quad \text{for } \gamma_v \in (\tilde{\gamma}_v,\ \gamma_0), \\ &= (\gamma_{min} + \gamma_0)/2 \quad \text{if } \overline{\alpha} \neq \alpha_{IV}(V^*(\gamma)) \quad \text{for } \gamma \in (0,\ \gamma_0) \end{aligned} \tag{19}$$

where $\tilde{\gamma}_v \equiv max\{0,\ \gamma_{**}\}$ and we define $\gamma_{min} = 0$ for $\overline{\alpha} = \alpha_{IV}(Z_1)$.

The upper bound of α in the interval (10) is estimated by using the formula (8) for $\widehat{\overline{\alpha}}(p)$ and procedures explained in Appendix A, with $Q = \{Z_1, V_1, V_2, V^*(\gamma)\}$ and $\gamma = \hat{\gamma}$ defined in (19). Although an estimator of the lower bound $\underline{\alpha}$ can be defined as $\widehat{\underline{\alpha}}(p)$ in a similar way to $\widehat{\overline{\alpha}}(p)$, the only candidate for $\underline{\alpha}$ is $\alpha_{IV}(\omega(\gamma))$. For this reason, $\widehat{\underline{\alpha}}(0.50)$ is simply estimated by $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$ and $\widehat{\underline{\alpha}}(0.975)$ is estimated by the left end of a 95% (two-sided) confidence interval of $\alpha_{IV}(\omega(\hat{\gamma}))$. To estimate these bounds, we recall our previous discussion and choose IIV pairs (Z_1, Z_2) with and without the village mean of farm work hours, $C_5 = \bar{X}_{-j}$. Specifically, we choose pairs (C_5, C_4) and (C_1, C_4) for the eastern region and pairs (C_5, C_1) and (C_2, C_3) for the other regions, where $C_1 = num\_farmer$, $C_2 = num\_farmer\_male$, $C_3 = num\_farmer\_female$, and $C_4 = \bar{W}_{-j}$ (village mean of production costs).

Table 6 presents estimates of $\widehat{\overline{\alpha}}(0.50)$ and $\widehat{\underline{\alpha}} = \hat{\alpha}_{IV}(\omega(\hat{\gamma}))$, where estimates of $\widehat{\overline{\alpha}}(0.975)$ and the left ends of two-sided 95% confidence intervals of $\alpha_{IV}(\omega(\hat{\gamma}))$ are shown in brackets. It also presents estimates of $\hat{\alpha}_{IV}(Z_j)$, $\hat{\alpha}_{IV}(V_j)$, $\hat{\alpha}_{IV}\big(V^*(\hat{\gamma})\big)$, $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$, values of various constants $(\tau_{xyz},\ \gamma_{min},\ \gamma_0,\ \gamma_*,\ \gamma_{**},\ \hat{\gamma},\ \gamma_{NR})$, and estimated lower bounds $\hat{\alpha}_{IV}(\omega(\gamma))$ with $\gamma = 0.5$ and $\gamma = \gamma_{NR}$. The negative values of τ_{xyz} show that condition (C5) or (C5') is satisfied for all IIV pairs in each region.

The two IIV pairs produce very similar estimates of $\widehat{\overline{\alpha}}(0.50)$ in each region, although we find various values of the upper bound using a single IIV in Table 5. For the selected pairs, $\widehat{\overline{\alpha}}(0.50)$ is estimated at 0.321 and 0.322, 0.184 and 0.181, and 0.112 and 0.128 in the

eastern, central, and western regions, respectively. On the other hand, if we use C_1, C_4, or C_5 (C_1, C_2, C_3, or C_5) as a single IIV in the eastern region (in the other regions), Table 5 shows that $\widehat{\bar{\alpha}}(0.50)$ is estimated to be 0.32-0.33, 0.18-0.22, and 0.09-0.16 in the eastern, central, and western regions, respectively. Furthermore, the two IIV pairs estimate $\widehat{\bar{\alpha}}(0.975)$ to be 0.361 and 0.365, 0.185 and 0.201, and 0.118 and 0.161 in the eastern, central, and western regions, respectively. If we use relevant IIVs as a single IIV, Table 5 shows that $\widehat{\bar{\alpha}}(0.975)$ is estimated to be 0.39-0.56, 0.23-0.31, and 0.23-0.26 in the eastern, central, and western regions, respectively. These results suggest that paired IIVs produce a lower and more precise estimate of the upper bound of α (particularly $\widehat{\bar{\alpha}}(0.975)$) than a single IIV. They also suggest that choice of C_5 does not have a significant impact on estimates of the upper bound of α.

On the other hand, the two IIV pairs produce different estimates of $\underline{\hat{\alpha}}$ in each region. For IIV pairs with and without C_5, $\underline{\hat{\alpha}}$ is estimated at 0.201 and 0.143, 0.140 and 0.111, and 0.023 and 0.059 in the eastern, central, and western regions, respectively. Furthermore, these IIV pairs estimate the left end of a two-sided confidence interval of $\alpha_{IV}(\omega(\hat{\gamma}))$ to be 0.022 and -0.241, 0.002 and -0.408, and -0.132 and -0.863 in the three regions.

Since the lower bound $\alpha_{IV}(\omega(\hat{\gamma}))$ depends on values of γ, we explain these results by focusing on two values of γ. We first focus on γ_*, which gives one of the asymptotes of a rectangular hyperbola representing $\alpha_{IV}(\omega(\hat{\gamma}))$. From definition $\gamma_* = \sigma_{z_1\tilde{x}}/(\sigma_{z_1\tilde{x}} + \sigma_{z_2\tilde{x}})$, it is closely related to correlation between Z_1 and $\tilde{X}$. Our discussion in Table 4 finds that this correlation is the highest for $Z_1 = C_5$, which tends to raise γ_*, given correlation of $\tilde{X}$ with Z_2. A higher value of γ_* makes the asymptote $\gamma = \gamma_*$ to stand more rightward in Figure 1, which slows down the speed of decreasing $\alpha_{IV}(\omega(\gamma))$ as γ increases in the interval $(0, \gamma_*)$. Table 6 shows that values of γ_* for IIV pairs with and without C_5 are 0.74 and 0.48, 0.66 and 0.51, and 0.51 and 0.54 in the eastern, central, and western regions, implying that the slow-down effect related to γ_* is stronger in the eastern and central regions.

We next focus on $\hat{\gamma}$, which is determined in two ways for the results shown in Table 6. Specifically, $\hat{\gamma}$ is given by $\gamma_0/2$ for IIV pair (C_1, C_4) in the eastern region and (C_5, C_1) in the western region. For the other cases, $\hat{\gamma}$ is given by $(\gamma_{min} + \gamma_0)/2$. In the former case, $\hat{\gamma}$ is very small (0.09 or 0.07) because $\hat{\alpha}_{IV}(Z_1)$ is the smallest of the four candidates of the upper bound $\overline{\alpha}$ and therefore γ_0, which sets $\alpha_{IV}(\omega(\hat{\gamma}))$ to be zero, is small. In the latter case, $\hat{\gamma}$ ranges from 0.33 and 0.52. Given the position of $\gamma = \gamma_*$, a higher value of $\hat{\gamma}$ causes a lower value of $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$. Consequently, the case of $\hat{\gamma} = \gamma_0/2$ tends to produce a higher estimate of $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$, whereas the case of $\hat{\gamma} = (\gamma_{min} + \gamma_0)/2$ tends to produce a lower estimate of $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$.

These interpretations of γ_* and $\hat{\gamma}$ can be used to explain why IIV pairs with C_5 tend to raise estimates of the lower bound $\alpha_{IV}(\omega(\gamma))$ in all regions. In the eastern region, $\hat{\gamma}$ is higher but the slow-down effect related to γ_* is much stronger for pair (C_5, C_4). In the central region, the values of $\hat{\gamma}$ are similar but the slow-down effect is stronger for pair (C_5, C_1). In the western region, the slow-down effects are similar but the value of $\hat{\gamma}$ is much lower for pair (C_5, C_1). In particular, these relations explain much lower values of the left end of the confidence interval of $\alpha_{IV}(\omega(\hat{\gamma}))$: IIV pairs without C_5 estimate this left end at -0.24, -0.41, and -0.86 in the eastern, central, and western regions.

Now, using IIV pairs with C_5, we construct an interval of α from $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$ and $\widehat{\overline{\alpha}}(0.50)$ to compare it with the OLS estimate of α. This interval is estimated to be (0.20, 0.32), (0.14, 0.18), and (0.02, 0.11) in the eastern, central, and western regions, respectively, each of which does not include the corresponding OLS estimates 0.43, 0.20, and 0.16 in the three regions. Furthermore, we use the same IIV pairs to construct another interval of α from the left end of the confidence interval of $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$ and $\widehat{\overline{\alpha}}(0.975)$, which can be compared with a 95% confidence interval of α derived from OLS estimation. The former interval is estimated to be (0.02, 0.36), (0.00, 0.19), and (-0.13, 0.12) for the eastern, central, and western regions, respectively, each of which have no or narrow intersection with the latter

intervals (0.37, 0.48), (0.15, 0.25), and (0.08, 0.23) in the three regions. These results show serious bias of the OLS estimates and the related confidence intervals.

Finally, we compare the values of $\hat{\gamma}$ determined by (19) and those set by Nevo and Rosen (2012). Table 6 shows in most cases that $\hat{\gamma}$ is smaller than 0.5 and γ_{NR}, both of which are smaller than γ_*. This relation means that $\alpha_{IV}(\omega(\hat{\gamma}))$ is higher than $\alpha_{IV}(\omega(0.5))$ and $\alpha_{IV}(\omega(\gamma_{NR}))$. Furthermore, when we choose IIV pairs without C_5, γ_{NR} takes very close values to 0.5 and γ_* and causes extremely large positive or negative estimates of $\hat{\alpha}_{IV}(\omega(\gamma_{NR}))$. These results show that our choice of γ helps obtain a narrow interval of α particularly when we choose C_5 as an IIV.

Conclusion

This study adopts and improves the imperfect instrumental variables (IIV) method of Nevo and Rosen (2012) in production function estimation using cross sectional data. Because the error term in the production function (productivity shock) typically has a positive correlation with labor input, IIVs are also required to have positive correlation with the error term. Consequently, we must use two IIVs together to derive the upper and lower bounds of production elasticities.

To apply this method to cross sectional data of Chinese farm households, we choose a key IIV, the mean of farm work hours for other households in the village, and combine it with the other IIV, the number of farm workers or the mean of production costs for other households in the village. Furthermore, by finding important properties and relations of candidates for the upper and lower bounds of the production elasticity of labor, we propose an alternative way to determine the weight of combining two IIVs, which uses more information from data and economic theory.

Our estimation results show that OLS estimates of the production elasticity of labor are

not included in the corresponding intervals estimated from our IIV method. The results also show that 95% confidence intervals of the elasticity derived from OLS estimation only have narrow intersection with the corresponding "95% confidence intervals" estimated from our IIV method.

Appendix A: Detailed Procedures to Compute Principal Critical Values

This appendix explains procedures to compute the principal critical value $k_{\hat{Q}}(p)$ in the simulation-based estimator (8), by following the Algorithm 1 of Chernozhukov, Lee, and Rosen (2013). For notational readability, we denote variables Z and V in the text by 1 and 2, and write $\bar{\alpha} = min_{q\in Q}\{\alpha_{IV}(q)\}$, $Q = \{1,2\}$ in this appendix.

Define column vector $\boldsymbol{\alpha}_{IV} = (\alpha_{IV}(1), \alpha_{IV}(2))'$ and let $\boldsymbol{\Omega}$ denote the covariance matrix of the limiting distribution of $n^{1/2}(\widehat{\boldsymbol{\alpha}}_{IV} - \boldsymbol{\alpha}_{IV})$ (n: sample size), where $\widehat{\boldsymbol{\alpha}}_{IV}$ and $\widehat{\boldsymbol{\Omega}}$ respectively denote consistent estimators of $\boldsymbol{\alpha}_{IV}$ and $\boldsymbol{\Omega}$. Then, $k_{\hat{Q}}(p)$ is computed in five steps:

(Step 1) Obtain 400 estimates of $\widehat{\boldsymbol{\alpha}}_{IV}$ by bootstrap, use them to compute a sample covariance matrix $n^{-1}\widehat{\boldsymbol{\Omega}}$ of $\widehat{\boldsymbol{\alpha}}_{IV}$, and obtain a matrix $(n^{-1}\widehat{\boldsymbol{\Omega}})^{1/2} = (\widehat{\omega}^*_{qr})$ $(q, r = 1,2)$.

(Step 2) Define $\widehat{\boldsymbol{\omega}}^*(q) = (\widehat{\omega}^*_{q1}, \widehat{\omega}^*_{q2})$ and compute $(q) = \|\widehat{\boldsymbol{\omega}}^*(q)\|$ $(q = 1,2)$.

(Step 3) Generate standard normal random variables G_{1i} and G_{2i} $(i = 1, \ldots, 1000)$, which are statistically independent, and define $\boldsymbol{G}_i = (G_{i1}, G_{i2})'$.

(Step 4) Obtain $Z^*_i(q) = \widehat{\boldsymbol{\omega}}^*(q)\boldsymbol{G}_i/s(q)$ $(q = 1,2)$ for each i and define the auxiliary critical value $k_Q(\gamma_n)$ as the γ_nth quantile of $max_{q\in Q}\{Z^*_i(q)\}$, where $\gamma_n = 1 - 0.1/ln(n)$. This value is used to compute $\tilde{\alpha}_{IV} = min_{q\in Q}\{\hat{\alpha}_{IV}(q) + k_Q(\gamma_n)s(q)\}$ and to define the adaptive inequality selector $\hat{Q} = \{q \in Q \mid \hat{\alpha}_{IV}(q) \le \tilde{\alpha}_{IV} + 2k_Q(\gamma_n)s(q)\}$.

(Step 5) Use the same random sample of $Z^*_i(q)$ $(q = 1,2; i = 1, \cdots, 1000)$ in Step 4 but use the set $\hat{Q}$ to compute the pth quantile of $max_{q\in\hat{Q}}\{Z^*_i(q)\}$, which gives the principal critical value $k_{\hat{Q}}(p)$.

References

Blundell, R. and S. Bond (1999) "GMM Estimation with Persistent Panel Data: An Application to Production Functions" *Econometric Reviews* 19, pp. 321–340.

Chernozhukov, V., S. Lee, and A. M. Rosen. (2013). "Intersection Bounds: Estimation and Inference", *Econometrica* 81, pp. 667–737.

Griliches, Z. and J. Mairesse (1998) "Production Functions: The Search for Identification." In S. Strom ed. *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, Cambridge University Press.

Gustafsson, B. A., S. Li and T. Sicular (2008) *Inequality and Public Policy in China*. Cambridge University Press. New York.

Jacoby, H. G. (1993) "Shadow Wages and Peasant Family Labour Supply: An Econometric Application to the Peruvian Sierra." *Review of Economic Studies* 60, pp. 903–921.

Knight, J., Q. Deng, and S. Li. (2011) "The Puzzle of Migrant Labour Shortage and Rural Labour Surplus in China." *China Economic Review* 22, pp. 585–600.

Levinsohn, J. and A. Petrin (2003) "Estimating Production Functions Using Inputs to Control for Unobservables." *Review of Economic Studies* 70, pp. 317–341.

Li, S. (2002) *Chinese Household Income Project, 2002* [Computer file]. ICPSR21741-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2009-08-14. doi:10.3886/ICPSR21741.

Marschak, J. and W. Andrews (1944) "Random Simultaneous Equations and the Theory of Production." *Econometrica* 12, pp. 143–205.

Mundlak, Y. (1961) "Empirical Production Function Free of Management Bias". *Journal of Farm Economics* 43, pp. 44–56

Nevo, A. and A. M. Rosen. (2012) "Identification with Imperfect Instruments", *Review of*

*Economics and Statistic*s 94, pp. 659–671 (Working paper version was published in 2008 as cemmap working paper CWP16/08).

Olley, S. and A. Pakes (1996) "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64, pp. 1263–1298.

Stock, J. H. and M. Yogo (2005) "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, Cambridge University Press.

Yang, D. T. (1997a) "Education in Production: Measuring Labor Quality and Management." *American Journal of Agricultural Economics* 79, pp. 764–772.

Yang, D. T. (1997b) "Education and Off-Farm Work." *Economic Development and Cultural Change* 45, pp. 613–632.

Table 1. Sample Means of Variables by Region

Region	East		Center		West	
Sample size	1212		1642		1320	
Variables in production functions						
value_added [yuan]	4637	(5117)	4800	(4105)	4054	(3790)
farmlabor [hour]	2351	(1887)	2361	(1581)	3219	(1893)
capital [yuan]	4639	(15978)	3726	(6013)	3630	(6118)
land [mu]	6.399	(7.272)	8.800	(9.076)	6.825	(6.496)
irr_share	0.645	(0.398)	0.532	(0.424)	0.513	(0.367)
educ_head	7.723	(2.405)	7.343	(2.369)	6.761	(2.597)
age_head	47.31	(9.755)	43.73	(9.697)	44.57	(10.25)
Instrumental variables						
tapwater	0.420	(0.494)	0.164	(0.371)	0.283	(0.450)
firewood	0.658	(0.475)	0.576	(0.494)	0.695	(0.461)
ownland_share	0.969	(0.113)	0.979	(0.083)	0.977	(0.087)
num_adult_male	1.370	(0.579)	1.380	(0.591)	1.527	(0.725)
num_adult_female	1.330	(0.565)	1.344	(0.582)	1.404	(0.627)
num_childyt6	0.111	(0.325)	0.180	(0.414)	0.223	(0.482)
num_childover6	0.581	(0.738)	0.812	(0.812)	0.916	(0.932)
vill_large	0.288	(0.453)	0.312	(0.464)	0.361	(0.481)
vill_wage [yuan/day]	19.12	(6.307)	17.45	(4.872)	14.85	(3.651)
vill_p_rice [yuan/kg]	1.066	(0.098)	1.039	(0.189)	0.991	(0.099)
Imperfect instrumental variables						
pcost_all [yuan]	3298	(8768)	2517	(2789)	3315	(4505)
num_farmer_male	1.325	(0.544)	1.334	(0.562)	1.507	(0.711)
num_farmer_female	1.278	(0.521)	1.293	(0.541)	1.386	(0.619)

Note: Standard deviations are shown in parentheses and units are shown in brackets. One yuan is approximately equal to 0.16 dollar and one mu is approximately equal to 0.16 acre or 0.067 ha.

Table 2. OLS Estimates of Parameters of Production Function (1)

Region	East	Center	West
Sample size	1212	1642	1320
ln(farmlabor)	0.429*	0.201*	0.157*
	(0.028)	(0.025)	(0.037)
ln(capital)	0.023*	0.054*	0.077*
	(0.008)	(0.007)	(0.012)
ln(land)	0.471*	0.438*	0.438*
	(0.035)	(0.028)	(0.040)
irr_share	0.390*	0.078	0.626*
	(0.073)	(0.045)	(0.062)
educ_head	-0.001	0.006	0.012
	(0.010)	(0.007)	(0.009)
age_head	0.044*	0.026	0.019
	(0.020)	(0.014)	(0.016)
age_head^2	-0.005*	-0.003	-0.002
	(0.002)	(0.002)	(0.002)
Adjusted R^2	0.423	0.371	0.316

Note: Standard errors are shown in parentheses. The coefficient and standard error of age_head^2 are multiplied by 10. * indicates statistical significance at 5% level. Coefficients of province dummy variables are omitted to save space.

Table 3. IV Estimates of Parameters of Production Function (1)

Region	East		Center		West	
Sample size	1212		1642		1320	
IV set	Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
ln(farmlabor)	0.378*	0.485*	0.167	0.189*	0.347*	0.184
	(0.137)	(0.106)	(0.116)	(0.070)	(0.143)	(0.116)
ln(capital)	0.025*	0.020*	0.056*	0.055*	0.067*	0.076*
	(0.010)	(0.009)	(0.009)	(0.007)	(0.014)	(0.014)
ln(land)	0.488*	0.453*	0.446*	0.441*	0.384*	0.431*
	(0.056)	(0.048)	(0.038)	(0.032)	(0.056)	(0.050)
irr_share	0.397*	0.383*	0.073	0.076	0.636*	0.627*
	(0.074)	(0.074)	(0.047)	(0.045)	(0.063)	(0.062)
educ_head	-0.003	0.002	0.005	0.006	0.017	0.013
	(0.011)	(0.011)	(0.008)	(0.007)	(0.010)	(0.009)
age_head	0.047*	0.041*	0.028	0.027	0.013	0.018
	(0.021)	(0.020)	(0.016)	(0.015)	(0.016)	(0.016)
age_head^2	-0.005*	-0.004*	-0.003	-0.003	-0.002	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$\bar{R}^2$	0.275	0.290	0.280	0.337	0.347	0.365
F stat.	50.49	8.586	78.88	23.57	94.65	14.48
	{16.38}	{38.54}	{16.38}	{38.54}	{16.38}	{38.54}
OIR stat.	NA	32.27	NA	19.41	NA	17.10
		[0.000]		[0.022]		[0.047]
D-W-H stat.	0.147	0.294	0.091	0.034	1.922	0.059
	[0.701]	[0.588]	[0.763]	[0.853]	[0.166]	[0.809]

Note: IV set 1 includes only the number of adults in the household. IV set 2 includes the ten instrumental variables listed in Table 1. Standard errors are shown in parentheses ( ). The coefficient and standard error of age_head^2 are multiplied by 10. * indicates statistical significance at 5% level. $\bar{R}^2$ and F stat. are the adjusted coefficient of determination and F statistic to test joint significance of instruments in the first stage regression, respectively. The number in braces { } is a critical value of the Wald statistic given by Stock and Yogo (2005), which tests significance of *ln(farmlabor)* in production function (1) at the 5% level. OIR stat. is χ^2 statistic for overidentifying restrictions test, which has degrees of freedom equal to (number of IVs) – 1, and p-values are shown in brackets [ ]. D-W-H stat. is Durbin-Wu-Hausman statistic to test endogeneity of *ln(farmlabor)*. Coefficients of province dummy variables are omitted to save space.

Table 4. Correlation of $\tilde{X}$ with candidates for imperfect instrumental variables (IIVs)

Region	East		Center		West	
Sample size	1212		1642		1320	
IIV candidate Z	hh vars	v. mean	hh vars	v. mean	hh vars	v. mean
num_farmer	0.164	0.050	0.207	0.115	0.240	0.034
	(5.79)	(1.74)	(8.58)	(4.69)	(8.99)	(1.24)
num_farmer_male	0.141	0.024	0.170	0.104	0.195	0.043
	(4.96)	(0.82)	(6.97)	(4.25)	(7.21)	(1.54)
num_farmer_female	0.122	0.063	0.168	0.091	0.195	0.015
	(4.29)	(2.20)	(6.90)	(3.69)	(7.20)	(0.55)
$ln(pcost\_all)$	0.236	0.167	0.183	0.105	0.153	0.046
	(8.43)	(5.91)	(7.56)	(4.26)	(5.60)	(1.68)
$ln(farmlabor)$	0.864	0.567	0.866	0.591	0.833	0.499
	(59.80)	(23.95)	(69.97)	(29.64)	(54.58)	(20.90)

Note: Column "hh vars" (household-level variables) shows correlation $\rho_{z\tilde{x}}$ between $\tilde{X}$ and IIV candidate Z, which equals $\rho_{x\tilde{x}}$ for $Z = ln(farmlabor)$. Column "v. mean" shows similar correlation between $\tilde{X}$ and the village-mean of Z, where the mean is taken for data of others than household j in the village. The number in parentheses shows t-value for the coefficient of Z in the regression of $\tilde{X}$ on Z.

Table 5. Upper Bounds of the Production Elasticity of Labor (α) Estimated with a Single IIV

IIV (Z)	C_1	C_2	C_3	C_4	C_5
Eastern region: $\hat{\alpha}_{OLS}$ = 0.429 (0.028), sample size = 1212					
$\hat{\alpha}_{IV}(Z)$	0.232	0.346	0.094	0.957	0.492
	(0.171)	(0.195)	(0.238)	(0.187)	(0.049)
$\hat{\alpha}_{IV}(V)$	0.475	0.445	0.484	0.302	0.309
	(0.048)	(0.047)	(0.047)	(0.043)	(0.070)
$\widehat{\bar{\alpha}}(0.50)$	0.320	0.450	0.223	0.323	0.329
	[0.558]	[0.526]	[0.570]	[0.389]	[0.397]
β	-0.234	-0.195	-0.165	-0.240	-1.907
Central region: $\hat{\alpha}_{OLS}$ = 0.201 (0.025), sample size = 1642					
$\hat{\alpha}_{IV}(Z)$	0.295	0.229	0.364	1.279	0.182
	(0.121)	(0.147)	(0.150)	(0.349)	(0.042)
$\hat{\alpha}_{IV}(V)$	0.171	0.194	0.162	0.053	0.241
	(0.044)	(0.043)	(0.043)	(0.040)	(0.069)
$\widehat{\bar{\alpha}}(0.50)$	0.188	0.210	0.178	0.072	0.219
	[0.241]	[0.262]	[0.230]	[0.132]	[0.314]
β	-0.315	-0.244	-0.241	-0.137	-2.149
Western region: $\hat{\alpha}_{OLS}$ = 0.157 (0.037), sample size = 1320					
$\hat{\alpha}_{IV}(Z)$	0.287	0.233	0.348	1.566	0.041
	(0.154)	(0.189)	(0.191)	(1.155)	(0.074)
$\hat{\alpha}_{IV}(V)$	0.104	0.134	0.098	0.074	0.331
	(0.069)	(0.067)	(0.069)	(0.058)	(0.088)
$\widehat{\bar{\alpha}}(0.50)$	0.135	0.164	0.129	0.107	0.089
	[0.232]	[0.258]	[0.227]	[0.211]	[0.262]
β	-0.406	-0.305	-0.305	-0.059	-1.495

Note: Imperfect instrumental variables (IIVs) include $C_1 = num\_farmer$, $C_2 = num\_farmer\_male$, $C_3 = num\_farmer\_female$, $C_4 = \bar{W}_{-j}$ and $C_5 = \bar{X}_{-j}$, where $\bar{W}_{-j}$ and $\bar{X}_{-j}$ denote sample means of $ln(pcost\_all)$ and $ln(farmlabor)$ for others than household j in the village. $\hat{\alpha}_{OLS}$ and $\hat{\alpha}_{IV}(Z)$ respectively denote OLS and IV estimators of α and their standard errors are shown in parentheses, where $V = \sigma_x Z - \sigma_z X$. $\widehat{\bar{\alpha}}(0.50)$ denotes a half-median unbiased estimator of the upper bound of α and the right end of 97.5% confidence interval of this bound is shown in brackets. β is the weight to express $\alpha_{IV}(V)$ as a linear combination of α_{OLS} and $\alpha_{IV}(Z)$ in equation (6).

Table 6. Upper and Lower Bounds of the Production Elasticity of Labor (α) Estimated with Paired IIVs

	East		Center		West	
Sample size	1212		1642		1320	
IIVs (Z_1, Z_2)	(C_5, C_4)	(C_1, C_4)	(C_5, C_1)	(C_2, C_3)	(C_5, C_1)	(C_2, C_3)
$\underline{\hat{\alpha}} = \hat{\alpha}_{IV}(\omega(\hat{\gamma}))$	0.201	0.143	0.140	0.111	0.023	0.059
	[0.022]	[-0.241]	[0.002]	[-0.408]	[-0.132]	[-0.863]
$\hat{\overline{\alpha}}(0.50)$	0.321	0.322	0.184	0.181	0.112	0.128
	[0.361]	[0.365]	[0.185]	[0.201]	[0.118]	[0.161]
$\hat{\alpha}_{IV}(Z_1)$	0.492	0.232	0.182	0.229	0.041	0.233
$\hat{\alpha}_{IV}(Z_2)$	0.957	0.957	0.295	0.364	0.287	0.348
$\hat{\alpha}_{IV}(V_1)$	0.309	0.476	0.241	0.194	0.331	0.134
$\hat{\alpha}_{IV}(V_2)$	0.302	0.302	0.173	0.162	0.104	0.098
$\hat{\alpha}_{IV}(V^*(\hat{\gamma}))$	0.369	0.388	0.183	0.192	0.109	0.148
γ_{**}	0.251	-0.510	-0.603	0.155	-17.409	0.313
γ_{min}	0.449	0.000	0.145	0.260	0.000	0.382
γ_0	0.591	0.180	0.541	0.398	0.130	0.434
γ_*	0.738	0.475	0.656	0.512	0.514	0.535
$\hat{\gamma}$	0.520	0.090	0.343	0.329	0.065	0.408
τ_{xyz}	-0.020	-0.010	-0.003	-0.001	-0.005	-0.001
γ_{NR}	0.454	0.480	0.401	0.501	0.337	0.535
$\hat{\alpha}_{IV}(\omega(0.5))$	0.236	7.712	0.058	-2.463	-4.403	-0.541
$\hat{\alpha}_{IV}(\omega(\gamma_{NR}))$	0.297	36.721	0.122	-13.590	-0.188	-162.060

Note: Imperfect instrumental variables (IIVs) include $C_1 = num\_farmer$, $C_2 = num\_farmer\_male$, $C_3 = num\_farmer\_female$, $C_4 = \bar{W}_{-j}$ and $C_5 = \bar{X}_{-j}$, where $\bar{W}_{-j}$ and $\bar{X}_{-j}$ denote sample means of $ln(pcost\_all)$ and $ln(farmlabor)$ for others than household j in the village. $\hat{\alpha}_{IV}(Z)$ denotes an IV estimator of α with Z used as an instrument, where $\omega(\gamma) = \gamma Z_2 - (1-\gamma)Z_1$, $V_j = \sigma_x Z_j - \sigma_{z_j} X$, and $V^*(\gamma) = \sigma_x \omega(\gamma) - \sigma_{\omega(\gamma)} X$. $\hat{\overline{\alpha}}(0.50)$ denotes a half-median unbiased estimator of the upper bound of α and the right end of 97.5% confidence interval of this bound is shown in brackets. The lower bound $\underline{\alpha}$ is given by a point estimate of $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$ and the left end of a 95% confidence interval of $\hat{\alpha}_{IV}(\omega(\hat{\gamma}))$ is shown in brackets. τ_{xyz} is a scalar defined in condition (C5) and various values of γ (γ_{min}, γ_0, γ_*, γ_{**}, $\hat{\gamma}$, γ_{NR}) are defined in Figure 1.

Figure 1. Relation among Various Estimators of Production Elasticity of Labor (α)