Levy Economics Institute of Bard College

Working Paper No. 1036

The Estimation of Production Functions with Monetary Values

Jesus Felipe
School of Economics, De La Salle University

John McCombie
Cambridge University

and

Aashish Mehta
University of California Santa Barbara

January 2024

We are grateful to Javier Birchenall, Zach Bleemer and Asha Sundaram for helpful conversations. The usual disclaimer applies.

Contact: Jesus Felipe (jesus.felipe@dlsu.edu.ph), John McCombie (jslm2@cam.ac.uk), Aashish Mehta (mehta@global.ucsb.edu)

Levy Economics Institute
P.O. Box 5000
Annandale-on-Hudson, NY 12504-5000
http://www.levyinstitute.org

ISSN 1547-366X

ABSTRACT: For decades, the literature on the estimation of production functions has focused on the elimination of endogeneity biases through different estimation procedures to obtain the correct factor elasticities and other relevant parameters. Theoretical discussions of the problem correctly assume that production functions are relationships among physical inputs and output. However, in practice, they are most often estimated using deflated monetary values for output (value added or gross output) and capital. This introduces two additional problems—an errors-in-variables problem, and a tendency to recover the factor shares in value added instead of their elasticities. The latter problem derives from the fact that the series used are linked through the accounting identity that links value added to the sum of the wage bill and profits. Using simulated data from a cross-sectional Cobb-Douglas production function in physical terms from which we generate the corresponding series in monetary values, we show that the coefficients of labor and capital derived from the monetary series will be (a) biased relative to the elasticities by simultaneity and by the error that results from proxying physical output and capital with their monetary values; and (b) biased relative to the factor shares in value added as a result of a peculiar form of omitted variables bias. We show what these biases are and conclude that estimates of production functions obtained using monetary values are likely to be closer to the factor shares than to the factor elasticities. An alternative simulation that does not assume the existence of a physical production function confirms that estimates from the value data series will converge to the factor shares when cross-sectional variation in the factor prices is small. This is, again, the result of the fact that the estimated relationship is an approximation to the distributional accounting identity.

KEYWORDS: Endogeneity; Monetary Values; Physical Quantities; Production Functions

JEL CLASSIFICATIONS: C18; C81; C82

1. INTRODUCTION

Economists have estimated production functions for almost a century. The main objective therein is to obtain estimates of the factors' output elasticities, and of their elasticities of substitution. The estimates are used to test a wide variety of hypotheses, calculate total factor productivity, build economic models (growth), and inform policy. And yet, the profession has long been concerned with the possibility that estimates of OLS regressions may be possibly biased by endogeneity problems. Such problems arise because producers decide their factor inputs in response to information—not available to the econometrician—about their production possibilities (Marschak and Andrews 1944).

This has led to a plethora of techniques attempting to solve the endogeneity problem using different estimators.[1] Given that different estimators produce different parameter estimates, a key question is: which of these estimates are the correct ones?

In attempting to decide, researchers often settle for the estimates that make the most sense. Van Beveren (2012, 99) argues that, "in practice evaluation of factor elasticities is more problematic due to the absence of a definite prior on how high (or low) factor elasticities *should* be [italics in the original]." However, guided by neoclassical theory, researchers often prefer estimates that closely resemble the shares of value added received by each factor, as these can be easily interpreted. In the words of Solow (1974, 121, emphasis added), "when someone claims that production functions work, he means (a) that they give a good fit to input-output data without the intervention of the factor shares and (b) *that the function so fitted has partial derivatives that closely mimic the observed factor shares*." This assessment continues to describe common practice today. Indeed, estimated factor elasticities that deviate significantly from the factor

[1] Some examples are as follows (this is far from an exhaustive review of methods): Mundlack (1961), panel data with fixed effects; Hall (1988), use of firms' first-order conditions; Olley and Pakes (1996), allow for input endogeneity with respect to a time-varying unobservable: invert an input demand function to control for unobserved productivity shock; Blundell and Bond (2000), a GMM estimator (a lagged-input instrumental variables); Levinsohn and Petrin (2003) (a variant of Olley and Pakes' approach), invert an intermediate input demand function to control for unobserved productivity shock; Ackerberg, Caves, and Frazer (2015), similar to Olley and Pakes and Levinshon and Petrin but inverting instead "conditional" rather than "unconditional" input demand functions to control for unobserved productivity; Gandhi, Navarro, and Rivers (2020), nonparametric identification strategy for gross output. The latest papers offer useful reviews of estimation methods, and Van Beveren (2012) provides a comprehensive survey in the context of the estimation of total factor productivity.

shares are quickly dubbed as unreasonable and biased, and are attributed to regressors' endogeneity or to using the wrong estimator.

Endogeneity is a legitimate concern that arises in the context of a production function specified as a relationship between the physical quantities of output (widgets) and inputs (labor, machines, intermediate material inputs). Indeed, it has been generally accepted since Cobb and Douglas (1928) that a production function is a technological relationship between *physical* inputs and *physical* output. Cobb and Douglas (ibid., 139) explicitly referred to the *volume of physical production*.

Yet, the reality is that, due to the lack of physical data, researchers use value (or monetary) data for output, the capital stock and intermediate inputs (employment is the only variable expressed as a physical quantity). The recent literature using firm-level data has started to acknowledge that some measure of deflated monetary values is used because physical data are not available.

Levinsohn and Petrin (2003, footnote 17) acknowledged that applied work uses monetary values, even plant-level data. The measures of output, capital, materials, electricity, and fuels, in their empirical exercise, are real monetary values (of the base year). Yet, they sidestepped the issue and proceeded as if all variables were physical quantities. Katayama, Lu, and Tybout (2009) and Van Beveren (2012) were also clear that the series used are in value terms, deflated values, rather than physical quantities. Ackerberg, Caves, and Frazer (2015, Section 4.3.4) provided a useful review of methods to deal with the endogeneity problem. They added a section on "units of measurement" where they stated that, the other series, aside from labor, used to estimate production functions are expressed in monetary units. Yet, their view seemed to be that the endogeneity problem could be dealt with under some conditions, e.g., when firms' prices are observed, one can certainly retrieve the physical units. This corresponds to our discussion of the errors-in-variables problem (see below). This is certainly true but these prices are generally not available, and the solution is valid only for firm-level data in physical quantities and their corresponding prices (which hardly exist), not for aggregate data (deflated monetary values and price deflators).

In empirical applications with value data, discussions of endogeneity have proceeded for decades *as if* the input and output data were physical quantities. The approach of the recent literature is to devise estimators that can address the endogeneity problem despite using monetary values.

In this paper, we show that the use of monetary values raises two additional problems in empirical exercises, independent of endogeneity. One is an errors-in-variables issue. This arises because the monetary values of capital and output are imperfect proxies for physical quantities. It leads to coefficient estimates that are biased relative to the factor elasticities. This is acknowledged but perhaps underappreciated in the literature (De Loecker 2011; Van Beveren 2012; Collard-Wexler and De Loecker 2016).

The other problem is that the inputs and output used are related through the distributional accounting identity of output (in monetary terms) that has to hold in any properly collected dataset. This identity can be rewritten in a form that resembles a production function, even though it is not a production function because it is not a relationship in physical units. The identity argument is that the regression of output on inputs in monetary values will simply estimate the transformed accounting identity. As a consequence, the estimated coefficients of the factors will gravitate toward the average shares of each factor in the value of output. This would lead the researcher to conclude that production is subject to constant returns to scale and that factor prices equal their marginal-value products, no matter what the underlying production relations are (Phelps-Brown 1957; Samuelson, 1979; Simon 1979a, 1979b). This problem is not widely recognized in the profession.

In practice, these three problems appear together and obscure the discussion of endogeneity. The objective of this paper is to separate and clarify the three problems, and to understand the relationship between them. We will explain when and why estimation of a regression of output on inputs using series expressed in monetary values will yield coefficients biased relative to the elasticities of output with respect to labor and capital, and when and why those same estimates will be biased relative to the factor shares. The coefficient estimates will capture production relations when the endogeneity and errors in variable biases are small relative to the elasticities, and will capture distributional relations when the biases relative to the shares are small. We illustrate these arguments by simulating datasets that assume a physical production function of known form exists, and attempting to recover its parameters using the corresponding value data. To our knowledge, this paper is the first to offer a unified treatment of the three problems and explain their relation to the use of monetary data.

Though not identical, our simulations are in the spirit of those of Fisher (1971) and Fisher, Solow, and Kearl (1977), whose objectives were to understand why and when aggregate production functions work. The difference with respect to our simulations is that they focused on the aggregation problem, specifically why aggregate production functions appear to work despite that the aggregation conditions were violated. By being specific about the fact that aggregate series must be monetary values linked through the identity, our results complement theirs and clarify some issues.

Our simulations show that using data in monetary values to extract the true elasticities implicit in the relationship with physical data (i.e., the production function), works only under unrealistic scenarios. One such scenario assumes that there is no endogeneity problem and that the prices of output and machinery vary very little across firms. In this case, OLS works. Another scenario assumes that the variation in these prices is negligible, but there is endogeneity, and in this case, OLS is biased but IV still works in theory. In the far likelier scenario of even modest unobserved variation in input prices, IV can only work if the instruments (which must now solve both the errors in variables and endogeneity problems) are independent of these price variations. It is hard to see how this condition could be met unless one had instruments measured in physical units—in which case, one would bypass the errors in variables problem by using the physical instead of the monetary series. Each of these scenarios assumes that a physical production function of known form between inputs and outputs exists. As this is questionable, we also show that, when such stable physical production relationships do not exist, the estimates get closer to the factor shares as factor prices vary less and less. Finally, we show that these problems do not exist if the data are in physical units. The problem, in short, is not about the choice of estimator, but the nature of the data.

The rest of the paper is structured as follows. Section 2 describes the three problems—regressors' endogeneity, errors in variables, and the accounting identity—and places them in the context of the use of physical quantities and value data. Section 3 explains the structure of the simulations. Section 4 discusses the results. Section 5 concludes. Appendix A shows that the transformed value-added accounting identity resembles a Cobb-Douglas production function. Appendix B explains why the identity critique does not arise with quantity data. Appendix C provides the summary statistics. Appendices D and E provide our simulation code.

2. STATEMENT OF THE THREE PROBLEMS AND THE USE OF MONETARY VALUES

Following the discussion in the introduction and to avoid confusion, we will use the term "production function" only when referring to a relationship in physical terms. We assume that physical production is actually governed by a standard Cobb-Douglas function:

$$Q_i = AL_i^{\alpha}K_i^{\beta}e^{u_i} \qquad (1)$$

where Q is the quantity of homogeneous physical output (in widgets), L is homogeneous labor (in workers or hours), K is homogeneous physical capital (in number of machines), and u captures unmeasured productivity differences across observations. The subscript i indexes firms. This formulation ignores materials for simplicity, although these are, of course, necessary to produce output Q.

As is standard in the literature, this error term (u) is composed of two parts, both unobservable to the econometrician. One represents shocks to production or productivity not observable by the firm before makes its input decisions. It represents, for example, deviations from expected breakdown, defect, or rainfall amounts in a given year. The second part of u represents "productivity" shocks that are potentially observed or predictable by firms when they make input decisions. Intuitively, it represents variables such as the managerial ability of a firm, expected down-time due to machine breakdown, expected defect rates in a manufacturing process, soil quality, or the expected rainfall at a particular farm's location. It is this second part that gives rise to the endogeneity problem.

Regression (1) is typically estimated in logarithmic form. Denoting $lnL = l$ and $lnK = k$, the expected values of the estimates are:

$$E[\hat{\alpha}_{OLS}] = \alpha + E[\frac{\sigma_k^2\sigma_{l,u}-\sigma_{l,k}\sigma_{k,u}}{\sigma_l^2\sigma_k^2-(\sigma_{l,k})^2}] \quad (2a)$$

$$E[\hat{\beta}_{OLS}] = \beta + E[\frac{\sigma_l^2\sigma_{k,u}-\sigma_{l,k}\sigma_{l,u}}{\sigma_l^2\sigma_k^2-(\sigma_{l,k})^2}] \quad (2b)$$

The OLS estimator will be unbiased relative to the elasticities if log-labor (l) and log-capital (k) are exogenous with respect to output ($E[l,u] = E[k,u] = 0$), and biased otherwise.[2] This bias could undoubtedly be eliminated by using the correct estimation method, if the researcher had data on physical quantities of output (Q), labor (L), and capital (K), and knowledge of the production function (assume it is Cobb-Douglas).

It is worth noting that most researchers do not specify the units of the series used, and they simply refer to "output," "capital," and "labor." This point is important because, unlike in equation (1), the reality is that they have to use monetary values for output (denoted V —value added) and capital (denoted J —typically generated through the perpetual inventory method by adding investment) because there are no data on physical quantities. This means that, instead of equation (1), they effectively estimate:

[2] The general finding is that the elasticity of employment will be biased upward by endogeneity and that of capital downward. Van Beveren (2012) provides a comprehensive review of the literature of problems that affect the estimation of production functions (as well as of different estimators to deal with endogeneity in production functions). These problems include the bias due to the entry and exit of firms, the use of price deflators to proxy for firm-level prices, and the pervasiveness of multi-product firms in manufacturing.

$$V = AL^{\gamma}J^{\delta}e^{\varepsilon} \quad (3)$$

What happens then? How does the endogeneity problem come into play now? Is the solution to the potential endogeneity problem the same as with physical quantities in equation (1)? And what other problems beyond endogeneity are introduced? Stated differently, under what circumstances can the use of monetary values generate unbiased estimates for the correct elasticities: $E[\hat{\gamma}] = \alpha$ and $E[\hat{\delta}] = \beta$?

The use of monetary values for output (V instead of Q) and capital (J instead of K) introduces two problems. Both problems exist whether or not input levels are endogenously selected, so for clarity, the rest of this section assumes that there is no endogeneity bias.

The first problem involves errors in variables. The links (errors) between V and Q and between J and K, are the prices of output (P) and capital (P_K). For a single (homogenous) output and capital, these variables in physical units and in monetary values are linked by $V = PQ$, and $J = P_K K$. The production function (1) in physical terms can therefore be written in terms of monetary value variables as:

$$V = AL^{\alpha}J^{\beta}\left(\frac{P}{P_K^{\beta}}\right)e^{u} \quad (4)$$

If one had knowledge of P and P_K, and so of Q and K, it would make no difference whether the researcher estimated equation (1) or equation (4).[3] Both would yield the correct elasticities α and β. Therefore, so long as P and P_K are included as regressors, the use of monetary values would not pose problems.

The errors-in-variables problem appears when P and P_K are unknown or excluded. The researcher estimates equation (4), with the ratio of the two prices lumped into the error term.

[3] In the standard treatment of the errors-in-variables problem, as long as the error in the dependent variable (in our case, P, the result of using V instead of Q) and the regressors are uncorrelated, the OLS parameters are unbiased and consistent. However, the error in the dependent variable (in our case, P_K, the result of using J instead of K) leads to biased and inconsistent OLS estimates.

This creates an errors-in-variables bias if machine and widget prices are correlated with labor and value capital—as they will be, most obviously because firms using more expensive machines will obviously have a more valuable capital stock. If, for simplicity, we assume that $E[ln\ P\ , \varepsilon] = 0$, then using $\rho = -\beta\ ln\ (P_k)$ to simplify notation, the errors-in-variables bias is:

$$E[\hat{\gamma}_{OLS}] = \alpha + H = \alpha + E[\frac{\sigma_j^2 \sigma_{l,\rho} - \sigma_{j,\rho}\sigma_{l,j}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}] \quad (5a)$$

$$E[\hat{\delta}_{OLS}] = \beta + \Omega = \beta + E[\frac{\sigma_l^2 \sigma_{j,\rho} - \sigma_{l,\rho}\sigma_{l,j}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}] \quad (5b)$$

Three points about the errors-in-variables biases warrant emphasis. First, they are derived assuming there is no endogeneity bias. In reality, the total bias relative to the elasticities combines the effects picked up in equations (5a)–(5b) and (2a)–(2b).[4] We exclude an equation for this total bias for simplicity, tackling each of the three problems one at a time. Second, the errors in variables biases can be signed depending upon one's priors regarding the correlations that drive them, per equations (5a)–(5b). Finally, the errors-in-variables problem in this case is akin to one of omitted variables, except that, unlike in the standard omitted-variables problem, we know exactly which variables need to be included in order to recover the elasticities: $ln\ P$ and $ln\ P_k$.

The endogeneity and errors in variables problems are, in reality, compounded by the fact that the researcher might not know the true production function, and the fact that the data available might refer to some aggregates of output and capital, so there is no Q (single output, number of widgets) and K (single capital stock, number of homogeneous machines) in the statistics. All the researcher has is V and J, both aggregates in monetary values. The situation is worse if we think that different firms use different production processes with different production functions.

We move now to the third problem. Phelps-Brown (1957), Simon and Levy (1963), Simon (1979a, b), Samuelson (1979), and Shaikh (1980), among others, argued that the problem with

[4] For an expression of endogeneity bias in a value terms regression, simply replace k with j in (3).

estimations of equations like (3), is that the estimated coefficients of labor and capital will gravitate toward the factor shares in total value added.[5] The reason, they showed, is that the series used to estimate them, output, labor, and capital, are related through the accounting identity:

$$V \equiv W + \Pi \equiv wL + rJ \qquad (6)$$

where V is constant-price value added, W is the total wage bill, Π denotes all profits, w is the wage rate, and r is the profit rate. Note that the series V, W, Π are monetary values (and so are the products wL and rJ). Also note that identity (6) holds at any level of aggregation, firm level, sector of the economy, and the total economy (in this case, V would be GDP). [6]

Identity (6) can be rewritten as $V = Af(L,J)$. For example, for a cross-section of i firms with fairly constant factor shares, Appendix A shows that the identity (6) can be approximated by:

$$V_i \simeq exp(B)\, w_i^{\bar{a}} r_i^{1-\bar{a}} L_i^{\bar{a}} J_i^{1-\bar{a}} \qquad (7)$$

where $B = (ln\bar{V} - \bar{a}ln\bar{w} - (1-\bar{a})ln\bar{r} - \bar{a}ln\bar{L} - (1-\bar{a})ln\bar{J})$, and $\bar{a} \simeq (\bar{w}\bar{L})/\bar{V}$ and $(1-\bar{a}) \simeq (\bar{r}\bar{J})/\bar{V}$ are the average (denoted by a bar over the variable) labor and capital shares,

[5] Econometricians like Cramer (1969, 234–37), Intriligator (1978, 270), and Wallis (1979, 62–63), also went over algebra of the identity. Felipe and McCombie (2013) offer an extensive and unified treatment of the identity problem.

[6] A clarification: in the standard neoclassical set up total costs (TC) are $TC = wL + \varpi J$, where ϖ is the rental price, or user cost, of capital. The relation between this definition and the value-added identity equation (4) is $V \equiv TC + \Gamma$, where Γ denotes extra profits, which are part of total profits (Π) accrued to the owners of capital. These could also be written as the product of a profit rate (r') times the capital stock, i.e., $\Gamma = r'J$. This implies, logically, that $\Pi = rJ = \varpi J + \Gamma = (\varpi + r')J$; and so $r = \varpi + r'$. In the words of Samuelson (1979, 932; using our notation): "No one can stop us from labeling this last vector [residually computed profit returns to 'property' or to the nonlabor factor] as (rJ), as J. B. Clark's model would permit -even though we have no warrant for believing that noncompetitive industries have a common profit rate r and use leets capital (J) in proportion to the ($V - wL$) elements!" He also remarked that the accounting identity (6) is not the result of Euler's theorem (Samuelson 1979, 933). In other words, the identity must hold.

respectively. The weighted average of the factor prices $M_i = w_i^{\bar{a}} r_i^{1-\bar{a}}$ plays a very important role in the discussion of the identity argument. There are two important points about expression (7). First, it is no more than an approximation to the identity (not a true equation). It will have a high fit and its coefficients will approximate the factor shares if the variation in factor prices within the cross section is not great, which is often the case in empirical applications. In this case, M_i and $exp(B)\, w_i^{\bar{a}} r_i^{1-\bar{a}}$ will be fairly constant. Therefore, $V_i \simeq \tilde{A} L_i^{\bar{a}} J_i^{1-\bar{a}}$ (where $\tilde{A} = exp(B)\, w_i^{\bar{a}} r_i^{1-\bar{a}}$ will still be an approximation to the accounting identity (6). The obvious point is that when this is estimated as $V_i = A L_i^{\gamma} J_i^{\delta} e_i^{\varepsilon}$ it is indistinguishable from equation (3). The second important point is that it resembles the Cobb-Douglas function. It is indeed the same functional form, but here in monetary values.[7] Other transformations of the accounting identity are certainly possible. They would look like CES or translog functions, for example.

As above, the question is: when a researcher estimates this equation, is he/she estimating a production function or the accounting identity?[8]

Equations (5a)–(5b) show that the estimates will be biased relative to the elasticities by the omission of P and P_k from equation (4). An analogous derivation shows that the same estimates will also be biased relative to the average factor shares by the omission of M from equation (7):

$$E[\hat{\gamma}_{OLS}] = \bar{a} + \Psi = \bar{a} + E\left[\frac{\sigma_j^2 \sigma_{l,\mu} - \sigma_{j,\mu}\sigma_{l,j}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}\right] \tag{8a}$$

$$E[\hat{\delta}_{OLS}] = (1-\bar{a}) + \Phi = (1-\bar{a}) + E\left[\frac{\sigma_l^2 \sigma_{j,\mu} - \sigma_{l,\mu}\sigma_{l,j}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}\right] \tag{8b}$$

where $\mu = \ln M = \bar{a} lnw + (1-\bar{a}) lnr$. These arise due to covariation between $ln\, L$, $ln\, J$, and the omitted but known term $ln\, M$. The bias is smaller when $ln\, M$ varies less.

[7] Douglas (1976) did not seem to be aware of the Phelps-Brown (1957) and Simon and Levy (1963) papers, as he was still convinced at the time that the results he presented (estimates of the labor and capital coefficients equal to the factor shares, hence adding up to one. See Tables 1–5 in the paper) were meaningful.

[8] This argument applies equally to regressions estimated in growth rates as the identity holds also in growth rates.

The important point to note is that the first terms of the two expected values are the average corresponding factor shar. Moreover, because Ψ and Φ are merely the outcome of omitting M, they would disappear if the researcher estimated equation (3) by correcting for $\ln M$. In this case, the parameter estimates would be unbiased for the average factor shares. This would also be the case if $\ln M$ were omitted but was constant across the observations, or if the researcher corrected for both $\ln w$ and $\ln r$ rather than $\ln M$.

Equations (8a)–(8b) capture a peculiar form of omitted-variable bias. We label it "peculiar" because we know what the omitted variable is, namely $M_i = w_i^{\bar{a}} r_i^{1-\bar{a}}$. Of course, researchers do not typically estimate a production function including w and r. However, in practice, they do add variables such as human capital. As should be obvious (and we will show), the better such variables proxy for w and r, the closer the results would get to those predicted by the identity argument—that is, the estimated coefficients would approximate the factor shares and the fit (R-squared) would approach one.

Note that implicit in this derivation is the recognition that the data are (must be) monetary units, not physical quantities. This problem, neglected in discussions of production function estimations, is fundamental. Moreover, the difference between Q and V is not simply the units but the fact that while the former (Q) is generated through the production process (under the assumption that there is a production function for Q), the latter (V) is generated by the accountants or the statisticians exactly as $W + \Pi$, which is also equal to $wL + rJ$. The same argument applies if the measure of output were gross output instead of value added. [9] Consequently, equation (3) relates output to inputs but it does so in monetary terms *through* the accounting identity, not the production process.

Getting back to where we started, it ought to be clear by now that when a researcher estimates equation (3), $V = AL^{\gamma} J^{\delta} e^{\varepsilon}$, he/she faces a combination of three problems (endogeneity, errors in

[9] The identity would be $Y = wL + rJ + p_x x$: where J, L, and x are capital (in value terms), labor, and the quantity of materials, respectively; r, w, and p_x are their respective prices, and Y is gross sales.

variables, the identity problem) that can lead to coefficient estimates that are biased relative to the elasticities. Endogeneity bias is conceived in the literature as a problem with the use of data in physical units. Yet, it could equally arise with output and capital measured in monetary units. The errors-in-variables bias arises because monetary values are used to proxy physical capital and output. The identity problem is present always but its consequences are more subtle and, for this reason, the issue has not been clearly understood by the profession. Moreover, the bias in the estimated coefficients that results from the data being related through an accounting identity is of very different nature. In reality, all three issues—endogeneity, errors in variables and the identity—appear together, with the consequence that it is difficult to ascertain what causes the bias in the estimation of the elasticities. So, what exactly do estimates of equation (3) capture?[10]

The foregoing discussions set up a simple point, not previously noted in the literature: a single set of coefficient estimates on value data is simultaneously biased relative to the elasticities (by endogeneity or errors in variables) and relative to the average factor shares (by the omissions of w and r). It is possible to assess whether these estimates will better approximate the factors' shares in value added or the elasticities by understanding the likely size of these biases. If the endogeneity or errors-in-variables biases relative to the elasticities (equations [2] and [5]) are large and the omitted variables bias relative to the shares (equation [8]) is small, the estimates will be closer to the shares. If the converse is true, then they will be closer to the elasticities. This insight structures the thought experiments in the rest of this paper.

As the biases are not credibly measurable in real-world data, we estimate the biases in simulated datasets. We generate series in physical terms, as well as prices of output and capital that allow us to create the corresponding series in value terms. The average factor shares differ significantly from the elasticities in our datasets, allowing us to demonstrate which of the two the estimates capture under different conditions. Our analysis explains why, in most real-world settings, the

[10] It is of course possible that researchers believe the relationship among deflated monetary values is correct (i.e., it is a production function), and they are effectively interested in knowing such data as "the dollar increase in output when the capital stock increases by one dollar." If the objective is stated this way, then the errors-in-variables problem with production functions in value terms, is bypassed. However, the identity problem clearly remains. Moreover, if the object of interest really were the elasticity of *an additional dollar of capital*, then measurement errors due to unobserved variation in prices of capital could not be a problem—and the mainstream literature on the consequences of "unobserved errors in the prices, for both inputs and outputs" (Collard-Wexler and De Loecker 2016, 4) would have no reason to exist.

biases relative to the elasticities are likely to be large and those relative to the shares may be small. As a result, the estimates are more likely to reflect the distribution of value added among the factors, rather than production relations.

It is important to note that the problem appears much more simplified here than it would in reality. This is because of our choice of functional form, the Cobb-Douglas, which simplifies the simulations, and which has allowed us to conceptualize the errors-in-variables as one of omitted-variables biases. With a CES production function, firms using different production processes, or aggregation issues concerning output, inputs, and production functions (Felipe and Fisher 2003), the problem would become much more complex to deal with.

3. SIMULATED DATA

We simulate data for a cross-section of 100,000 hypothetical firms. Appendix C provides a table of summary statistics, while Appendix D provides our main simulation code, which proceeds as follows:

- The firms (indexed by i) share a Cobb-Douglas production function in physical quantities with firm-specific technical efficiency (d_i). The econometrician cannot observe d_i. All firms produce the same homogenous product, and, for simplicity, there are no intermediate materials.[11] Output in physical terms is determined by $Q_i = d_i L_i^{\alpha} K_i^{\beta}$, where $d = Ae^u$. Our benchmark parameterization assumes constant returns to scale, with $\alpha = 2/3$ and $\beta = 1/3$. We also simulate an increasing-returns-to-scale production function with $\alpha = 4/3$ and $\beta = 2/3$.

[11] Not adding intermediate materials, while conceptually very important (a production process requires materials to produce physical output), does not affect the essence of our simulations. With monetary values, equation (6) for gross output (instead of value added) would include the value of materials. The latter's share would appear when the expression is written in multiplicative form (like equation (7)) and the three shares (in gross output) would add up to one. Conceptually, adding materials to our simulations would help reinforce the point that researchers do not use physical quantities because these are not available, or because at any level of aggregation, all we have is values.

- Labor (L_i) and technical efficiency (d_i) are drawn from independent log-normal distributions $\ln(L) \sim N(\ln(20), 0.1)$ and $\ln(d) \sim N(\ln(50), \sigma_d)$, respectively.
- Physical capital, K_i, measures the number of machines of a homogenous type utilized by firm i. We generate two alternative series of physical capital and output. The first one assumes that firms select the machinery level without observing d_i and this physical capital series is generated as: $K_i^* = \exp(\ln R_i + lnL_i)$, where $\ln R \sim N(0, \sigma_R)$ is the log of the capital–labor ratio. This setup allows for some limited variation in the choice of technique (to allow for identification of the individual factor elasticities), with firms averaging one machine per worker. The second series, which is endogeneity-contaminated, is $K_i^e = \exp(lnK_i^* + \ln s_i)$, where $\ln s_i = [\ln d_i - \ln(50)] + v_i$ and $v_i \sim N(0, \sigma_v)$. This captures a scenario in which firms observe s_i—a mean-zero signal of d_i that is distorted by noise (v_i). Firms decide to use more machines if the signal is large. Two physical quantity output series (in widgets) are then generated, one exogenous, $Q_i^* = d_i L_i^\alpha K_i^{*\beta}$, and another one contaminated by endogeneity, $Q_i^e = d_i L_i^\alpha K_i^{e\beta}$.
- Under our baseline parameterization ($\sigma_d = 0.1, \sigma_\theta = 0.05, \sigma_v = 0.07$), the physical capital and output series are highly correlated across the datasets [$Corr(lnK^*, lnK^e) = 0.675,\ Corr(lnQ^*, \ln Q^e) = 0.980$]. Also, the endogeneity-uncontaminated capital and technical efficiency series are uncorrelated by construction ($Corr(lnd, lnK^*) = 0.003$). These conditions ensure that lnK^* will be a suitable instrument for lnK^e.

These physical quantity series will suffice to explore simultaneity biases, and solutions to them. We will also use them to demonstrate that the identity argument does not apply when estimating production functions from data in physical units (This is shown in Appendix B). To explore the problems that arise from working with data in monetary values, and what it would take to overcome them, we transform physical-quantity data into value terms:

- The purchase price per machine varies across firms with a dollar value $P_{K,i} \sim U[P_K^* - \Delta_k, P_K^* + \Delta_k]$, so that the values of the capital stocks with and without simultaneity are given by $J_i^* = P_{K,i}K_i^*$ and $J_i^e = P_{K,i}K_i^e$, respectively.[12]
- The price of widgets is distributed across firms as $P_i \sim U[P^* - \Delta_P, P^* + \Delta_P]$, so that the corresponding value-added series (value added equals revenue given the assumed absence of raw materials) are $V_i^* = P_i Q_i^*$ and $V_i^e = P_i Q_i^e$.
- These value terms output and capital series are imperfect proxies for their physical-quantities counterparts. We assume that $(P_K^*, P^*) = (\$200{,}000/machine, \$1{,}333/widget)$.
- To show what happens as we change the correlations between the data series in physical and value terms ($\rho(K^*, J^*)$ and $\rho(Q^*, V^*)$), we run our simulations under four sets of values for (Δ_k, Δ_P): "very high" $(\$1{,}000, \$13.3)$, "high" $(\$10{,}000, \$133)$, "moderate" (\$30,000, \$200), and "low" correlation (\$50,000, \$1,000). Under these assumptions, the average capital–output ratio $\left(\frac{J}{V}\right)$ works out to roughly three, a fairly common figure in value-terms datasets.
- Finally, to complete the series in monetary values, we simulate the factor incomes. No optimization condition is imposed. Our factor input series consciously violate the first-order condition that the value of each factor's marginal product equals its price. This condition is equivalent to assuming that the factor output elasticities are equal to their respective shares in output. Since a central objective of our simulations is to clarify when and why the regression coefficients provide biased estimates of the factor shares and the elasticities, this objective cannot be met if the two have the same value.

We assume that firm i rents labor for an annual wage rate w_i, where $w \sim U[w^* - \Delta_w, w^* + \Delta_w]$, with $(w^*, \Delta_w) = (\$16{,}667/\text{worker}, \$10{,}000/\text{worker})$, and that it earns an annual return on

[12] The price of capital is conceptualized differently in theory and in practice. In theory, K is a stock but output is a flow. Therefore, the assumption is made that the stock is a good proxy for the flow of capital services during the time period in physical units (e.g., square meters of office), while its price is measured as a rental rate (dollar per square meter per year). However, the datasets actually used to estimate production functions calculate capital as a stock in monetary values using the perpetual inventory method (i.e., they measure J in dollars of office space). Thus, P_K is the purchase price of a unit of physical capital (dollars per square meter), not the rental price per unit of time. As value capital is measured as a stock, not a flow, r must be an annual rate of return in order for $rJ = rP_kK$ to be measured in the correct units (dollars/year).

capital that is determined residually by the accounting identity, $r_i \equiv (V_i - w_i L_i)/J_i$. This set up ensures that the accounting identity always holds in value terms, as it must in any properly collected dataset. It will suffice, for our purposes, to generate the return on capital only in the exogenous capital case, so that $r_i \equiv (V_i^* - w_i L_i)/J_i^*$.

Having these factor incomes permits us to demonstrate the identity problem. As shown in equations (5a)–(5b) and Appendix A in detail, the coefficients of lnV^* on lnL and lnJ^* should approximate the factor shares if these are relatively constant in the cross section, and the regression corrects for $lnM_i = \bar{a} lnw_i + (1 - \bar{a}) lnr_i$ (this would not be needed if lnM_i is constant). To show that this works, we first create a normally distributed noise variable, z, with the same mean and standard deviation as $ln\ M_i$, and then use this to construct a series of proxies, indexed by $\phi \in [0,1]$, defined as $ln\ W_\phi \ = \phi * ln\ M \ + (1 - \phi)z$. We will show that, as $\phi \to 1$, the coefficients of a regression of lnV^* on $ln\ W_\phi$, $ln\ L$, and lnK^*, approach the factor shares.

Most importantly, irrespective of whether endogeneity is present or not, these parameter values ensure that the average capital share $(1 - \bar{a}) \equiv 1 - (\bar{w}\bar{L})/\bar{V}$ is roughly 0.75, which is starkly different from the benchmark output elasticity of capital, β =0.333. Thus, errors in variables will drive the estimated capital coefficient below its true elasticity of 0.333, while the identity problem will move the coefficient upwards from the elasticity of 0.333 toward the capital share of 0.75. Moreover, these simulated datasets can be used to populate equations (2), (5), and (8), in order to demonstrate why a given set of OLS coefficients estimated with value data provides biased estimates of both the elasticities and the shares.

Setting up the simulations this way will permit us to isolate each of the three estimation problems. We can turn the simultaneity bias on (off) by working with the endogeneity-contaminated (uncontaminated) physical quantities dataset. We can turn the classical errors in variables problem up and down by increasing and reducing Δ_k. And, finally, we can create an identity problem by increasing ϕ toward 1.

4. RESULTS

We now use these simulated data to analyze sequentially each of the three problems. We estimate OLS and 2SLS regressions of equations (1) and (3) as follows:

$$lnQ^{*} = ln\,A + \alpha lnL + \beta lnK^{*} + \boldsymbol{\tau}'\boldsymbol{X} + u \qquad (9a)$$

$$ln\,Q^{e} = ln\,A + \theta lnL + \pi\, ln\,K^{e} + \boldsymbol{\tau}'\boldsymbol{X} + u \qquad (9b)$$

$$ln\,V = ln\,\tilde{A} + \gamma lnL + \delta lnJ + \tilde{\boldsymbol{\tau}}'\boldsymbol{X} + \varepsilon \qquad (9c)$$

The regressions differ by whether or not they use exogenous-physical, endogenous-physical, or (exogenous) money-value, measures of output and capital; and by whether or not they include controls for ln P, ln P_k and ln W_ϕ, notated here simply as **X**. We will review the properties of the endogeneity-contaminated estimates $(\hat{\theta}, \hat{\pi})$, but then focus on the conditions under which the value-terms estimates $(\hat{\gamma}, \hat{\delta})$ approximate the output elasticities $(0.667\ for\ labor, 0.333\ for\ capital)$ and the shares (roughly 0.25 for labor, 0.75 for capital).

A. The Endogeneity Problem: Estimation in Physical Units

Table 1 shows the results of the production function estimated with data in physical units, corresponding to equations (9a) and (9b). The results confirm the standard understanding of the simultaneity problem and its solution *when working with data in physical units* (and an idealized production process whose functional form is known a priori). Panel (a) shows that OLS recovers the elasticities when all data are measured in physical units and capital is exogenously determined. Panel (b) shows, as expected, that OLS overestimates the output elasticity of capital and underestimates the output elasticity of labor when more productive firms choose to employ more machines. Panel (c) shows that 2SLS solves this problem, so long as a suitable physical-units instrument for endogenous physical capital exists. This confirms that the endogeneity problem is solvable, at least in principle, when working with data in physical units.[13]

[13] Griliches and Mairesse (1998, Section 6) consider the possibility of using factor prices to instrument for endogenous factors. This does not work in our simulated dataset because our need to ensure that a≠α guarantees that factor prices would be weak instruments for factor quantities. This is shown by our code in Appendix D.

Table 1: Estimation with Data in Physical Units

Dependent Variable	Coefficient on			R-Squared
(A) OLS: Capital and Output are not contaminated by endogeneity (9a)				
	lnK^*	lnL	Constant	
lnQ^*	0.344 (0.006)	0.656 (0.007)	3.911 (0.009)	0.5079
(B) OLS: Capital and Output are contaminated by endogeneity (9b)				
lnQ^e	lnK^e	lnL	Constant	
	0.905 (0.002)	0.098 (0.003)	3.903 (0.006)	0.8516
(C) 2SLS: lnK^*instrument for $ln\, ln\, K^e$ (9b)				
First Stage	lnK^*	lnL	Constant	
$ln\, K^e$	-0.020 (0.008)	1.015 (0.008)	0.013 (0.012)	0.4556
Second Stage	Predicted lnK^e	lnL	Constant	
$ln\, Q^e$	0.344 (0.006)	0.656 (0.007)	3.910 (0.009)	0.6585

Source: Authors

Table 2 implements equation (2) to decompose the endogeneity bias of the OLS estimates from Table 1, Panel (B). As expected, given that $ln\, L$ is uncorrelated with u in our data, these results indicate that the biases arise due to the correlation between $ln\, K^e$ and u ($\sigma_{k,u}$).

Table 2: Decomposition of the Simultaneity Bias (Data in Physical Quantities)

	Elasticity	**OLS Estimates**	**Bias**	**Decomposition of the Expected Bias**		
				$\frac{\sigma_k^2\sigma_{l,u}}{\sigma_l^2\sigma_k^2-(\sigma_{l,k})^2}$	$\frac{-\sigma_{k,u}\sigma_{l,k}}{\sigma_l^2\sigma_k^2-(\sigma_{l,k})^2}$	$Total$
Labor	0.667	0.098	-0.569	0.0008	-0.5700	-0.569
				$\frac{\sigma_l^2\sigma_{k,u}}{\sigma_l^2\sigma_k^2-(\sigma_{l,k})^2}$	$\frac{-\sigma_{l,u}\sigma_{l,k}}{\sigma_l^2\sigma_k^2-(\sigma_{l,k})^2}$	$Total$
Capital	0.333	0.905	0.572	0.5727	-0.0003	-0.572

Note: The empirical decomposition is calculated using the actual variances and covariances in the physical-quantities dataset. The OLS estimates are those in Table 1, Panel (B).
Source: Authors

B. Estimation in Value Terms 1: Errors in Variables

Table 3 provides coefficients from several versions of the regression when output and capital are value measures, equation (9c). To focus on the errors-in-variables problem introduced by the use of value data to proxy for physical quantities, we shut down all possibility of simultaneity by using the endogeneity-uncontaminated datasets. The errors-in-variables bias arises due to the use of a noisy value-terms proxy for lnK^*(i.e., due to estimation of [9c] instead of [9a]).

Notice that, just as (1) can be rewritten as (4), (9a) can be rewritten as:

$$lnV_i^* = (lnd_i + ln\,P_i\,) + \alpha ln\,L_i\ + \beta - ln\,P_{k,i}\,) + u_i \tag{10}$$

If the series for $ln\,ln\,P$ and $ln\,ln\,P_k$ are available, the OLS estimator will yield unbiased estimates of α and β.

We have assumed, for simplicity, that $corr(\ln P, \ln d) = 0$, so that the error introduced by replacing lnQ_i^* with lnV_i^*, is classical and its omission introduces no bias. In this case, the errors in variable biases due to replacing $\ln K^*$ with $\ln J^*$ are given by (5a) and (5b), and those equations permit us to predict the expected direction of the bias. Note that, by construction, $-\beta \ln\,(P_k)$ is negatively correlated with $\ln \bar{J}$ (i.e., $\sigma_{j,\rho} < 0$), but uncorrelated with $\ln L$ (i.e., $\sigma_{l,\rho} = 0$). We therefore expect from (5a) and (5b) a negative bias in the capital coefficient ($\Omega < 0$), and a positive bias in the labor coefficient ($H > 0$).[14]

This understanding of the effects of using a value proxy for physical capital has four corollaries:

C1. The estimated OLS coefficient of capital will be smaller than $\beta = 0.333$, and that of labor will be larger than $\alpha = 0.667$.

[14] It is also the case that $\sigma_{l,j} \approx \sigma_l^2$ in our simulated datasets, so that the scale elasticity is unbiased by the errors-in-variables problem ($H + \Omega \approx 0$). This need not be the case in general.

C2. The bias (relative to the elasticity) in the capital coefficient will become more negative as machine prices become more variable. This is expected because the negative covariance between machine values and prices ($\sigma_{j,\rho}$) becomes more pronounced the more machine prices vary (i.e., as Δ_k increases).

C3. When the regression is estimated in values correcting for both $\ln P$ and $\ln P_k$, the coefficient of $\ln P$ will be one, while the coefficient of $\ln P_k$ will be $-\beta = -0.333$. Moreover, the coefficients of labor and value capital will match the elasticities α and β. This is apparent from an examination of equation (10). Moreover, dropping $\ln P$ from the regression will not change this outcome.

C4. The IV regression in values, using $\ln K^*$ as an instrument for lnJ^* , will recover the elasticities. This is obvious, and relies on the imposed assumption that $\ln K^*$ is orthogonal to $\ln P_k$, the "error" in the independent variable $\ln J^*$.

Table 3 provides results that confirm these four corollaries. It presents the results of four value-terms regressions. Each is estimated on a sequence of four datasets, characterized by progressively lower correlations between the physical- and value-terms data (Panel A). The very high correlation case results are indistinguishable from those in physical units. C1 is apparent from every coefficient in Panel B other than the very high correlation case—the OLS-estimated capital coefficients are less than 0.333 and the labor coefficients exceed 0.667. C2 is confirmed by comparison across the columns of Panel B. All elements of C3 are confirmed by Panels C and D—all of the coefficients on $ln\ P$ are insignificantly different from 1; the coefficients on lnP_k all approximate –0.333, those on lnJ and lnL approximate the elasticities regardless of whether we control for lnP_i or not. Finally, the 2SLS estimates in Panel D approximate the elasticities, confirming C4.

Table 3: Errors in Variables Bias Relative to the Elasticities

		Very high correlation	High Correlation	Moderate Correlation	Low Correlation
		(1)	(2)	(3)	(4)
(A) Correlations between physical and value series					
	$Corr(K^*, J^*)$	0.9997	0.9683	0.7896	0.6073
	$Corr(Q^*, V^*)$	0.9992	0.9268	0.8534	0.2683
(B) OLS Regression of Value Added on Labor and Value Capital					
	Coefficients				
	log Labor	0.657 (0.007)	0.743 (0.007)	0.918 (0.006)	0.967 (0.020)
	log Value Capital	0..343 (0.006)	0.257 (0.006)	0.083 (0.004)	0.031 (0.011)
	R-squared	0.507	0.433	0.361	0.035
(C) OLS Regressions correcting for Machine and Output Prices					
	Coefficients				
	log Labor	0.656 (0.007)	0.656 (0.007))	0.656 (0.007)	0.656 (0.007)
	log Value Capital	0.344 (0.006)	0.344 (0.006)	0.344 (0.006)	0.344 (0.006)
	log Machine Price	-0.475 (0.110)	-0.357 (0.013)	-0.349 (0.007)	-0.347 (0.007)
	log Output Price	1.027 (0.055)	1.003 (0.005)	1.002 (0.004)	1.000 (0.001)
	R-squared	0.509	0.578	0.642	0.965
(D) OLS Regressions correcting for Machine Prices					
	Coefficients				
	log Labor	0.656 (0.007)	0.652 (0.008)	0.650 (0.009)	0.621 (0.037)
	log Value Capital	0.345 (0.006)	0.348 (0.007)	0.351 (0.008)	0.377 (0.033)
	log Machine Price	-0.479(0.110)	-0.365 (0.015)	-0.357 (0.010)	-0.387 (0.035)
	R-squared	0.507	0.436	0.370	0.0363
(E) 2SLS Regressions: Exogenous Physical Capital ($ln\ K^*$); Instruments for Exogenous Value Capital ($ln\ J^*$)					
	First stage Regression				
	Coefficients				
	log Labor	-0.000 (0.0002)	-0.002 (0.002)	-0.001 (0.006)	-0.001 (0.010)
	log Physical Capital	1.000 (0.0002)	1.000 (0.002)	1.001 (0.005)	1.001 (0.009)
	F (2, 99,9997)	7.53 x 10^7	752,371	82,783	29,221
	P-value	0.0000	0.0000	0.0000	0.000
	R-squared	0.9993	0.938	0.624	0.369
	Second Stage Regression				
	Coefficients				
	log Labor	0.656 (0.007)	0.652 (0.008)	0.651 (0.010)	0.621 (0.037)
	log Value Capital	0.345 (0.006)	0.348 (0.007)	0.350 (0.009)	0.377 (0.033)
	R-squared	0.507	0.432	0.335	0.025

Note: Regressions produced from data generated by the code appearing as Appendix D.

Source: Authors

The preceding results provide a framework to understand the errors in variables bias. While this bias is sometimes acknowledged in the literature (e.g., Ackerberg. Caves, and Frazer 2015), its importance is not widely reflected in applied work. Scholars who take this problem seriously suggest various partial solutions involving instruments for value capital based on past investments, trade shocks, or imperfect competition in output markets (De Loecker 2011; Van Beveren 2012; Collard-Wexler and De Loecker 2016). The availability of such partial fixes would provide comfort if the biases introduced by errors in variables were small. However, Panel A of Table 3 shows them to be extremely large. Even if $corr(\ln J, \ln K) = 0.97$, the estimated capital coefficient, 0.258, is 23 percent smaller than the elasticity (0.333). And with $corr(\ln J, \ln K) = 0.79$, the estimate, 0.082 is only a quarter of the elasticity.

Table 4 provides the decomposition of the expected EIV bias relative to the elasticities using equations (5a) and (5b). It shows that the bias is large because of the high covariance between the value of the capital stock and the price of machinery. It should be clear that this covariance will tend to be large in any dataset with somewhat variable machine prices, simply because $\ln J^* = \ln P_k + \ln K^*$. [15]

Table 4: Decomposition of the Errors in Variables Bias Relative to the Elasticities (High Correlation Dataset)

	Elasticity	OLS Estimates	Bias	Decomposition of the Expected Bias[a]		
				$\frac{\sigma_j^2 \sigma_{l,\rho}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}$	$\frac{-\sigma_{j,\rho} \sigma_{l,j}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}$	H
Labor	0.667	0.739	0.076	-0.0001	0.0827	0.082
				$\frac{\sigma_l^2 \sigma_{j,\rho}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}$	$\frac{-\sigma_{l,\rho} \sigma_{l,j}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}$	Ω
Capital	0.333	0.258	-0.075	-0.0827	0.0001	-0.082

Note: The empirical decomposition is calculated using the actual variances and covariances in the High Correlation dataset to proxy for their expected values. OLS estimates are from Column (2) of Table 3, Panel B.

Source: Authors

[15] Formally, $Cov(\ln J, \ln P_k) = Cov(\ln P_k + \ln K^*, \ln P_k) = V(P_k) + Cov(\ln K^*, \ln P_k)$, which will be large when machine prices vary greatly across firms (as they do), unless firms using many machines tend to buy them for substantially lower prices.

C. Estimation in Value Terms 2: The Identity Problem

We will now study the case in which the researcher estimates an approximation to equation (5), the accounting identity in multiplicative form. To do so, we shut down the errors-in-variables problem by focusing on the very high-correlation case, in which physical and value outputs are almost perfectly correlated.

Table 5 presents a series of estimates from regressions in value terms that correct for $ln\, W_{\phi}$. As expected, the baseline regression (Column 1), with $\phi = 0$, returns the elasticities, and not the shares. Columns (2)–(5) show clearly that the coefficient estimates converge on the factor shares as $ln\, W_{\phi}$ converges to $ln\, M$. Column (6) confirms that, when $ln\, W_1 = ln\, M$ is replaced with $ln\, w$ and $ln\, r$, the coefficients on these factor prices also return the factor shares. [16]

These results demonstrate that an econometrician using value data and hoping to recover the factor elasticities must take pains to exclude any control variables that could proxy for $ln\, M$. Including proxies for the factor prices ($ln\, M$), shifts the estimates toward the shares.[17]

[16] The reader will notice that the expression $M_i = exp(B)\, w_i^{\bar{a}} r_i^{1-\bar{a}}$ has to be identical to what the standard literature calls total factor productivity (TFP), typically calculated as $TFP_i = V_i/(L_i^{\bar{a}} J_i^{1-\bar{a}})$. This observation has two important corollaries: (i) $M \equiv TFP$ is always true, regardless of any theoretical restriction on producer behavior, cost functions, or whether the shares equal the elasticities, or not. While it is true that TFP is a weighted average of the factor prices, this is tautologically correct; and (ii) TFP derived from monetary values cannot be considered a measure of true productivity (in physical terms). As Appendix C shows, true TFP calculated with physical quantities (lnd) has a mean of 3.91 (or 49.89 for d), with a range of 3.51–4.33; while TFP calculated with monetary values (lmM^*) has a mean of 1.388 (or 4.00 for M^*), with a range of 0.701–1.891.

[17] We have used the Levinsohn and Petrin (2003) method with an actual data set (panel) that contains materials (required by this method). This data set is of course consistent with the gross-output accounting identity. As an experiment, we added the wage and profit rates as regressors in the first step. The coefficients estimated are the factor shares. Moreover, it makes no difference which variable, capital or labor, is the state variable. The estimated coefficients in both cases are the factor shares. See Felipe et al. (2021).

Table 5: Omitted Variables Bias Relative to the Shares (0.25 for labor, 0.75 for Capital)

OLS regressions that increasingly approximate the accounting identity						
	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline ($\phi = 0$)	$\phi = 0.25$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 1$	Full Accounting identity
Coefficients						
$ln\ L$	0.743 (0.0073)	0.698 (0.0070)	0.519 (0.0054)	0.293 (0.0018)	0.287 (0.0016)	0.284 (0.0015)
$ln\ J^*$	0.258 (0.0063)	0.303 (0.0060)	0.481 (0.0047)	0.707 (0.0016)	0.713 (0.0014)	0.716 (0.0013)
$ln\ W_\phi$	0.003 (0.0030)	0.360 (0.0036)	0.912 (0.0031)	1.026 (0.0008)	0.936 (0.0007)	
$ln\ w$						0.213 (0.0002)
$ln\ r$						0.704 (0.0005)
R-squared	0.433	0.485	0.696	0.966	0.973	0.976

Source: Authors

Table 6 returns to the case with no controls (equivalent to $\phi = 0$), but a high correlation between the physical- and value-terms series, and decomposes the omitted-variables bias relative to the shares, as in (8a) and (8b).

Table 6: Decomposition of the Omitted Variables Bias Relative to the Shares (High Correlation Dataset)

	Average Factor Share	**OLS Estimates**	**Bias**	**Decomposition of the Expected Bias**[a]		
				$\frac{\sigma_j^2 \sigma_{l,\mu}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}$	$\frac{-\sigma_{j,\mu} \sigma_{l,j}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}$	Ψ
Labor	0.252	0.743	0.491	0.0035	0.4836	0.4871
				$\frac{\sigma_l^2 \sigma_{j,\mu}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}$	$\frac{-\sigma_{l,\mu} \sigma_{l,j}}{\sigma_l^2 \sigma_j^2 - (\sigma_{l,j})^2}$	Φ
Capital	0.748	0.258	-0.490	-0.4838	-0.0026	-0.4864

Note: The empirical decomposition is calculated using the actual variances and covariances in the high correlation dataset to proxy for their expected values. OLS estimates are from Column (2) of Table 3, Panel B.

Source: Authors

A comparison of Tables 4 and 6, and equations (12a) and (12b) to equations (11a) and (11b) clarifies that the same set of OLS estimates on value data can suffer from two types of bias at the same time: errors in variables bias relative to the elasticities, and omitted variables bias relative to the shares. Whereas Table 4 showed that the errors in variables bias arises due to the covariance between $\ln P_k$ and $\ln J$, Table 6 shows that the omitted variables bias is driven principally by the covariance between $\ln M$ and $\ln J$.

Table 7 puts the problem in the starkest possible terms, showing that approximating the identity preempts estimates of the returns to scale, biasing the sum of the coefficients toward 1 (i.e., to the sum of the factor shares). Like Table 5, this table presents a series of OLS regressions that do an increasingly good job at correcting for $ln\, M$, but the underlying physical quantities dataset is regenerated with a scale-elasticity of 2 ($\alpha = 1.333,\ \beta = 0.667$). This does result in an unrealistically low average labor share ($\bar{a}$= 0.012), but saves on further and unnecessary recalibration of parameters. We present results for the high correlation case ($\Delta_k = \$10{,}000, \Delta_P = \$133$). The closer one gets to estimating the accounting identity, the closer the estimates become to the shares, and the closer the sum of the coefficients on labor and capital becomes to 1, despite a scale elasticity of 2. This explains why empirical estimates of "production functions" tend not to demonstrate evidence supporting the increasing returns to scale central to theories of endogenous growth and of trade. We find evidence of significant increasing returns to scale in columns (1), (2), and (3) (and only in the first two cases in the proximity of 2, corresponding to the case where value and physical terms are poorly correlated) but with estimates of coefficients of labor and capital that researchers would surely question.

Table 7: Omitted-Variable Bias Relative to the Shares (0.012 for labor, 0.988 for capital) when $\alpha + \beta = 2$

OLS regressions that increasingly approximate the accounting identity						
	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline ($\phi = 0$)	$\phi = 0.25$	$\phi = 0.5$	$\phi = 0.9$	$\phi = 1$	Full Accounting identity
Coefficients						
$ln\ L$	1.492 (0.0074)	1.404 (0.0072)	0.959 (0.0063)	0.045 (0.0012)	0.013 (0.0001)	0.013 (0.0001)
$ln\ J$	0.509 (0.0064)	0.538 (0.0067)	0.683 (0.0052)	0.977 (0.0009)	0.987 (0.0001)	0.987 (0.0001)
$ln\ W_{\phi}$	0.002 (0.0024)	0.236 (0.0030)	0..718 (0.0030)	1.087 (0.0005)	1.000 (0.0000)	
$ln\ w$						0.011 (0.0000)
$ln\ r$						0.987 (0.0000)
R-squared	0.750	0.765	0.840	0.995	1.000	1.000
Estimated "returns to scale"	2.000 (0.004)	1.942 (0.004)	1.641 (0.003)	1.022 (0.001)	1.000 (0.0000)	1.000 (0.0000)

Note: Data are from the High Correlation case.
Source: Authors

The takeaway from these three sets of results is that even if there is a well-defined production function with a known form (here Cobb-Douglas), econometricians hoping to estimate its parameters when using value-terms data must (i) solve the endogeneity problem, (ii) include controls for firm-specific machine prices, and (iii) exclude any control variables that proxy for the wage rate and the rate of return on value-capital. Instrumental variables can solve the endogeneity problem if the instruments for physical capital are uncorrelated with machine prices and productivity, but the most commonly used instruments (e.g., money values of lagged investment levels and trade shocks) are not likely to be so. Moreover, including variables that could proxy for firm-specific productivity or machine prices in order to solve problems (i) and (ii) will exacerbate problem (iii) to the extent that these proxies are correlated with $\ln w$ and $\ln r$.

D. Estimation in Monetary Terms if There Is No Production Function

This refers to the case when no properly specified production function between the observable physical variables exists. This may arise when there is more than one output, labor, or type of machine, and so the only way to express output and capital is in value terms. As is well known, an aggregate production function with neoclassical properties does not exist, in the sense that the conditions to aggregate output and inputs, as well as the production functions, are very hard to be satisfied by actual economies, and so its parameters cannot exist either (Felipe and Fisher 2003).

What might we expect in this case that the coefficients from regression of $ln\,V$ on $ln\,L$ and $ln\,J$ will capture? We argue that they would capture the only conceptual relationship that does exist between the observed variables: the transformed accounting identity (7). This argument is the same for firm-level data and for aggregate data. In this latter case, identity (6) would be the result of arithmetically summing up the identities for each firm or sector.[18]

To examine this possibility, we next generate a series of datasets that only exist in terms of money values and quantities of labor. There are no machines or widgets, and so no production function, but value added is distributed between labor and capital (equation [6]). Each dataset includes 100,000 firms, and contain labor (L), distributed as before, the labor share ($a \sim U[0.7,0.8]$), wages ($w \sim U[\$16,667 - \Delta_w, \$16,667 + \Delta_w]$), and returns to value capital ($r \sim U[0.10 - \Delta_r, 0.10 + \Delta_r]$). Each firm's value added is generated as $V = wL/a$, and its value capital as $J = (V - wL)/r$. The datasets are ordered by Δ_w and Δ_r, and therefore by the variation in $ln\,M$, permitting us to see what happens as this variation get smaller. The code is provided as Appendix E.

Unsurprisingly, in a regression of lnV on $\ln J$ and $\ln L$, the recovered coefficients begin far from the average factor shares (respectively, 0.25 and 0.75) when the variation in factor prices is large, and approach the average factor shares as that variation shrinks. For example, when $(\Delta_w, \Delta_r) = (\$1{,}000, 0.05)$ the estimated coefficients are 0.94 (on $\ln L$) and 0.06 (on $\ln J$), but they fall to 0.752 and 0.248 when $(\Delta_w, \Delta_r) = (\$100, 0.001)$.

It follows that when the functional form being estimated with value terms data corresponds closely to the transformed accounting identity, the regression will estimate the factor shares.[19]

[18] In this case, the aggregation of the corresponding identities yields: $\Sigma V_i = \bar{w}\Sigma L_i + \bar{r}\Sigma J_i = \bar{w}L + \bar{r}J$, where $\bar{w}$ and $\bar{r}$ are the average wage and profit rates, $\Sigma L_i = L$, and $\Sigma J_i = J$.

[19] These simulations complement those of Fisher (1971) using time series data (his concern was the aggregation problem). He generated the (aggregate) series of output, labor, and capital, by ensuring that the accounting identity held. Though Fisher did not state it, it is obvious that his aggregate series had to be monetary values, as it could not

There is no solution to this problem, and therefore it is not a matter of finding the correct estimator.

5. CONCLUSIONS

This paper has discussed three problems that econometricians face when they estimate production functions. Our departure point was the well-known problem caused by regressors' endogeneity, which leads to biased OLS estimates. For decades, the profession has proposed different estimators to solve this problem. The treatment of this problem assumes that output and capital stock are physical quantities (as they should be). We have shown that because, in practice, researchers use monetary values for output and capital, they face two additional problems that blur the treatment of endogeneity. One is a problem of errors in variables caused by the approximation of physical output and physical capital through monetary values. The other results from the fact that the series of output and inputs used (in monetary terms) are related though an accounting identity. The use of monetary (value) data is at the center of the problem.

We used simulation analyses to study if and when the coefficients obtained with the functions estimated with monetary data are in fact the true elasticities (corresponding to those of the production function, i.e., with physical quantities). We have assumed for our baseline discussions that there is an underlying Cobb-Douglas production function linking inputs and output, with constant returns to scale. This is an extremely important point as, in general, the researcher does not know what the functional form is.

Our simulations yield the following findings:

1. Estimation in physical units:

---

be otherwise. Fisher concluded that the Cobb-Douglas form works well (i.e., coefficients are close to the factor shares and the regression has a high fit) as long as factor shares are constant. Fisher stressed this last point and remarked that the constancy of the factor shares is not due to the fact that the underlying production function (of the economy) is Cobb-Douglas -it is the other way around: the Cobb-Douglas works if factor shares happen to be constant. Our results show that it is the variation in the weighted (by the factor shares) average of the wage and profit rate that determines how well the regression will estimate the factor shares.

    a. We corroborate that the estimation of production functions using physical data yields the correct factor elasticities if there is no endogeneity problem. If there is an endogeneity problem, an appropriate estimation method can solve it, but requires knowledge of the form of the production function and having the correct instruments.

2. Estimation using monetary values:
    a. *Errors-in-variables problem:* If the researcher controls for the prices of output and capital, estimation using value data is equivalent to the estimation using physical quantities. In case the regression does not control for these prices but the researcher has an appropriate instrument for the value of machines, instrumental-variable estimation will also recover the factor elasticities. However, such an instrument must be exogenous to the omitted machine prices, and so *most probably* cannot itself be in value terms. If such an instrument is not available, the use of value data will lead to biased coefficients with respect to the elasticities.
    b. *Accounting identity problem:* Including variables that correctly proxy for the weighted average of the factor prices in the regression will push the estimates toward the factor shares. If the value data are *very* highly correlated with the physical data (and so the previous finding on errors in variables is irrelevant), the regression can produce coefficients equal to the elasticities when there are no controls that proxy for the weighted average of the factor prices.
    c. *Accounting identity problem 2:* The use of value data will yield estimates of the elasticity of scale that are biased toward 1 (the sum of the factor shares) when the regression controls proxy for the weighted average of the factor prices. Tests of constant versus increasing returns to scale are therefore unreliable when using value data.
3. If there is no production function in physical quantities, all researchers can estimate with monetary data is an approximation to the distributional accounting identity. Coefficients will tend to approximate the factor shares if the weighted average of the wage and profit rate is correctly approximated, or if it varies little (in which case it is part of the constant term). If this weighted average varies substantially and it is not approximated through another variable, the estimated coefficients of labor and capital will deviate significantly

from the factor shares. This is most likely the problem many researchers face when they use aggregate data, and the problem the profession at large has tried to deal with for decades, without realizing. It has no econometric solution.

These results have implications for the significant body of applied literature that has been published during the last eight decades. They affect both aggregate exercises on growth and productivity and also exercises using firm-level data, to the extent that they all use monetary values.

Fisher, Solow, and Kearl (1977, 319) concluded that (aggregate) production functions should be used with extreme caution, "the way the old garbage man tells the young garbage man to handle garbage wrapped in plastic bags of unknown provenance: 'Gingerly, Hector, gingerly.'" Our simulations allow us to go one step further and advise that production function estimates from value data (at all levels of aggregation) should be left untouched.

REFERENCES

Ackerberg, D. A., K Caves, and G Frazer. 2015. “Identification Properties of Recent Production Function Estimators.” *Econometrica* 83 (6): 2411–51

Blundell, R. and S. Bond. 2000. "GMM estimation with persistent panel data: an application to production functions." *Econometric Reviews* 19(3): 321-340.

Cobb, C. and P. Douglas. 1928. “A Theory of Production.” *The American Economic Review* 18 (1): 139–65.

Collard-Wexler, A., and J. De Loecker. 2016. "Production function estimation with measurement error in inputs.”

Cramer, J. S. 1969. *Empirical Econometrics.* Amsterdam. North-Holland.

De Loecker, J. 2011. "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity." *Econometrica* 79: 1407–51.

Douglas, P. 1976. "The Cobb-Douglas Production Function once again: Its History, Its Testing, and Some New Empirical Values." *Journal of Political Economy*, 84(5): 903–16.

Felipe, J. and F. M. Fisher. 2003. “Aggregation in Production Functions: What Applied Economists Should Know.” *Metroeconomica* 54 (2–3): 208–62.

Felipe, J. and J. S. L. McCombie. 2013. *The aggregate production function and the measurement of technical progress: “Not even wrong.”* Cheltenham, UK and Northampton, MA: Edward Elgar.

Felipe, J., J. S. L. McCombie, A. Mehta, and D. F. Bajaro. 2021. “Production Function Estimation: Biased Coefficients and Endogenous Regressors, or a Case of Collective Amnesia?” Working Paper No. 994 (October). Levy Economics Institute of Bard College, New York.

Fisher, F. M. 1971. "Aggregate Production Functions and the Explanation of Wages: A Simulation Experiment." *Review of Economics and Statistics*, 52 (4): 305–25

Fisher, F. M., Solow, R. M., and Kearl, J. M. 1977. "Aggregate Production Functions: Some CES Experiments." *Review of Economic Studies* 44(2): 305–20.

Gandhi, A., S. Navarro, and D. A. Rivers. 2020. “On the Identification of Gross Output Production Functions.” *Journal of Political Economy* 128(8): 2973–3016.

Griliches, Z. and J. Mairesse. 1998. “Production Functions: The Search for Identification.” In *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch*

Centennial Symposium, edited by S. Strøm, pp.169–203. Cambridge, U.K.: Cambridge University Press.

Hall, R. 1988. "The Relation between Price and Marginal Cost in U.S. Industry." *Journal of Political Economy*, 96(5): 921–47.

Intriligator, M. D. 1978. *Econometric Models. Techniques and Applications.* Englewoods Cliffs. NJ: Prentice-Hall.

Katayama, H., S. Lu, and J. R. Tybout. 2009. "Firm-level productivity studies: illusions and a solution." *International Journal of Industrial Organization* 27: 403–13.

Levinsohn, J. and A. Petrin. 2003. "Estimating Production Functions Using Inputs to Control for Unobservables." *The Review of Economic Studies* 70 (2): 317–41.

Marschak, J. and W. H. Andrews. 1944. "Random Simultaneous Equations and the Theory of Production." *Econometrica* 12 (3/4): 143–205.

Mundlak, Y. 1961. "Empirical production function free of management bias." *Journal of Farm Economics* 43(1): 44–56.

Olley, G. S. and A. Pakes. 1996. "The Dynamics of Productivity in the Telecommunications Equipment Industry." *Econometrica* 64 (6): 1263–97.

Phelps-Brown, E. H. 1957. "The Meaning of the Fitted Cobb-Douglas Function." *The Quarterly Journal of Economics* 71(4): 546–60.

Samuelson, P. 1979. "Paul Douglas's Measurement of Production Functions and Marginal Productivities." *Journal of Political Economy* 87 (5): 923–39.

Shaikh, A. 1980. "Laws of Production and Laws of Algebra: Humbug II." In *Growth, Profits and Property*, edited by Edward J. Nell, 80–95. Cambridge: Cambridge University Press.

Simon, H. 1979a. "On Parsimonious Explanations of Production Relations." *The Scandinavian Journal of Economics* 81 (4): 459–71.

———. 1979b. "Rational Decision Making in Business Organizations." *The American Economic Review* 69 (4): 493–513.

Simon, H. A. and F. K. Levy. 1963. "A Note on the Cobb-Douglas Function." *Review of Economic Studies* 30(2): 93–94.

Solow, R. M. 1974. "Law of production and laws of algebra: the humbug production function: a comment." *The Review of Economics and Statistics* 56(1): 121.

Van Beveren, I. 2012. "Total factor productivity estimation: A practical review." *Journal of Economic Surveys* 26: 98–128.

Wallis, K. F. 1979. Topics in Applied Econometrics. The University of Minnesota Press. Minneapolis, Minnesota.

APPENDIX A: THE ACCOUNTING IDENTITY AND CROSS-SECTIONAL DATA

Factor shares in value added in a cross section can be written as $a_i = (w_i \, L_i)/V_i$ (labor) and $(1 - a_i) = (r_i \, J_i)/V_i$ (capital), where the subscript i refers to the cross-sectional unit. These expressions, when added up, lead to the accounting identity (for a cross section) $V_i \equiv W_i + P_i \equiv w_i L_i + r_i J_i$. For a low dispersion of the factor shares, $\bar{a} \simeq (\overline{w} \overline{L})/\overline{V}$, where a bar denotes the average value of a variable. Then, the following also holds:

$$a_i/\bar{a} \simeq [(w_i L_i)/V_i]/[(\overline{w}\overline{L})/\overline{V}] \qquad \text{(A1)}$$

and

$$(1 - a_i)/(1 - \bar{a}) \simeq [(r_i J_i)/V_i]/[(\bar{r}\bar{J})/\overline{V}] \qquad \text{(A2)}$$

For small deviations of a variable X_i from its mean $\overline{X}$, it follows that $ln\,(X_i/\overline{X}) \simeq (X_i/\overline{X}) - 1$. Taking logarithms of the previous two equations and using this approximation, it follows that:

$$ln\,[w_i/\overline{w}] + ln\,[L_i/\overline{L}] - ln\,[V_i/\overline{V}] \simeq (a_i/\bar{a}) - 1 \qquad \text{(A3)}$$

and

$$ln\,[r_i/\bar{r}] + ln\,[J_i/\bar{J}] - ln\,[V_i/\overline{V}] \simeq [(1 - a_i)/(1 - \bar{a})] - 1 \qquad \text{(A4)}$$

Multiplying equations (A3) and (A4) by $\bar{a}$ and $(1 - \bar{a})$, respectively, adding them, and rearranging the result yields:

$$\begin{aligned} lnV_i &\simeq B + \bar{a}lnw_i + (1 - \bar{a})lnr_i + \bar{a}lnL_i + (1 - \bar{a})lnJ_i = \\ &= B + lnM_i + \bar{a}lnL_i + (1 - \bar{a})lnJ_i \qquad \text{(A5)} \end{aligned}$$

where $B = (ln\bar{V} - \bar{a}ln\bar{w} - (1-\bar{a})ln\bar{r} - \bar{a}ln\bar{L} - (1-\bar{a})ln\bar{J})$ and $lnM_i = \bar{a}lnw_i + (1-\bar{a})lnr_i$. The weighted average of the factor prices lnM_i plays a very important role in the discussion of the identity argument. Taking anti-logs in (A5):

$$V_i \simeq exp(B)\, w_i^{\bar{a}} r_i^{1-\bar{a}} L_i^{\bar{a}} J_i^{1-\bar{a}} \qquad \text{(A6)}$$

Equation (A6) is an approximation to the accounting identity, and it will work well *if* the variation in factor prices within the cross section is not great, which is often the case in many empirical applications.

APPENDIX B: PRODUCTION FUNCTIONS WITH PHYSICAL QUANTITIES

The identity problem does not hold with physical data ($Q,\ L, K$) because these series are not linked through an accounting identity like the one with value (monetary) data. For simplification, we continue without adding materials, although the production function has to include the quantity of materials. With physical quantities, it would be possible to estimate the production function and obtain the true elasticities, which may differ from the factor shares.

Suppose there exists a production process that converts L (number of workers), K (number of identical machines) into Q (widgets). Assume that this production process is given by $Q = AL^{\alpha}K^{\beta}e^{u}$. This specific form is not essential to the argument. As this is an "engineering" or physical production function, output must be determined by the correctly measured flow of services from labor and capital.

Now a researcher estimates $lnQ = d + b_1 lnL + b_2 lnK + u$. What would the estimates of b_1 and b_2 pick up? We argue that she would obtain the true technological relationship, i.e., α and β (with an appropriate estimator, as discussed in the main text). In this case, the series Q, L, K are not definitionally related through an *actual* accounting identity.

Notice, though, that one could construct an infinite number of accounting identities with an *arbitrarily* chosen weight, b, that would determine the distribution of factor rewards in physical terms, e.g., $v = b(Q/L)$, $x = (1-b)(Q/K)$ with $0 < b < 1$, and then construct $Q \equiv vL + xK$, i.e., the identity in physical terms. As above, this expression could then be transformed into one in multiplicative form, the Cobb-Douglas production function. Note that v and x are measured in widgets per worker and widgets per machine, respectively. The important point now is that there is no *actual* identity relating Q, L, K. It is for this reason that the regression will pick up the true elasticities and not the factor shares—an infinite number, depending on the values of b.

However, with value data, the series for output (V), inputs (L, J), and the factor shares $a = wL/V$ and $(1-a) = rJ/V$, are related through only one identity. This is the identity that the monetary data regression will undoubtedly pick up.

Finally, it must be noted that, with physical quantities, researchers face two problems. First, is that one would need to know what functional form to estimate. The second problem is that, in practice, estimating a production function for manufacturing (or services) with physical quantities is next to impossible. This is due to the data requirements needed (e.g., individual capital stocks for an oil refinery). If these two issues were addressed, then the endogeneity of the regressors would be a correct concern because there is no identity directly linking Q, L, and K, and the regression would contain a true econometric error.

Appendix C: Summary Statistics

The following table summarizes our main dataset, with figures for value-terms variables drawn from the "high correlation" case.

Variable	**Mean**	**Std. Dev.**	**Min**	**Max**
(A) Variables in physical quantities				
$\ln d$	3.91	0.10	3.51	4.33
$\ln L$	3.00	0.10	2.57	3.42
$\ln K^*$	3.00	0.11	2.48	3.51
$\ln K^e$	3.00	0.17	2.33	3.72
$\ln s$	0.00	0.12	-0.50	0.53
$\ln Q^*$	6.91	0.14	6.32	7.60
$\ln Q^e$	6.91	0.17	6.20	7.73
(B) Variables in Value Terms				
P	1,333	77	1,200	1,466
P_k	199,996	5,768	190,000	210,000
w	16,665	5,754	6,667	26,666
r^*	0.252	0.050	0.083	0.498
J^*	4,023,648	466,527	243,813	6,762,812
V^*	1,346,956	208,403	685,466	2,508,228
a^*	0.252	0.093	0.070	0.618
$\ln M^*$	1.388	0.124	0.701	1.891

Appendix D: Stata Code for Sections 4.A-4.C

** NOTE TO USER: This code will create all the datasets and numerical results reported in this paper, other than those in Section 4.D (which are generated in Appendix E). To display the specific results you would like to verify, reset the relevant block of code from "quietly" to "noisily," check the assumed elasticities in lines 13–14, and set the relevant scenario in line 59.

```
clear all
set obs 100000
set seed 18

** Generate the data in physical units.
{
** We will use five global parameters to control the generation of physical quantities data.
global SD_lnR 0.05  /*Standard deviation of ln(machines per worker).*/
global SD_v 0.07 /*The standard deviation of the z in the technical efficiency signal. When it is zero, endogeneity bias is worst.*/
global SD_lnd = 0.1 /*Standard deviation of technical efficiency.*/
gen alpha = 0.667 /*Replace with 1.333 for IRS.*/
gen beta = 0.333 /*Replace with 0.667 for IRS.*/
** Generate the data.
gen lnL = rnormal(ln(20),0.1) /*Random labor input. */
gen L = exp(lnL)
gen lnd = rnormal(ln(50),$SD_lnd) /*Random technical efficiency parameter. */
gen d = exp(lnd)
gen lnR = rnormal(0,$SD_lnR) /*Random K/L ratio, centered on 1.*/
gen lnK_star = lnR + lnL /*Random capital level chosen by firm with no knowledge of its productivity parameter. */
gen K_star = exp(lnK_star)
gen lns = lnd - ln(50)+rnormal(0,$SD_v) /*Firms have a noisy signal of their technical efficiency. */
sum lnd lns /*Check the properties of the signal.*/
```

```
pwcorr lnd lns /*Check the properties of the signal.*/
gen lnK_end = lnK_star + lns /*Firms choose capital in response to this signal.*/
gen K_end = exp(lnK_end)
gen lnQ_star = lnd+alpha*lnL+beta*lnK_star /*ln(Output), if machines are determined
exogenously.*/
gen lnQ_end = lnd+alpha*lnL+beta*lnK_end /*ln(Output) if machines are determined
endogenously. */
gen Q_star = exp(lnQ_star)
gen Q_end = exp(lnQ_end)

** Examine the data.
pwcorr lnd lnL lnK_star lnK_end lnQ_star lnQ_end
sum
}

** RUN THE REGRESSIONS IN PHYSICAL QUANTITIES [Table 1]
{
** OLS recovers the elasticities when the data are not endogeneity-contaminated. [Panel (a)]
regress lnQ_star lnK_star lnL
lincom lnK_star + lnL

** OLS overestimates the capital elasticity when the data are contaminated by endogeneity.
[Panel (b)]
regress lnQ_end lnK_end lnL
lincom lnK_end + lnL

** K_star is a valid instrument for K_end, and 2SLS returns the elasticities. [Panel (c)]
pwcorr lnK_star lnK_end, sig
ivregress 2sls lnQ_end (lnK_end=lnK_star) lnL , first
lincom lnK_end + lnL
```

```
** Produce the stats used to calculate the decomposition of endogeneity bias [Table 2]
corr lnL lnK_end lnd, covariance
}

** MAKE THE VALUE-TERMS DATASETS AND RUN THE VALUE REGRESSIONS
** The prices of widgets and machines are uniformly distributed around means of $1,333 per
widget and $200,000 per machine. Here we set their half ranges to yield different levels of
correlation between physical and value capital/output.
foreach Correlations in High {/* Pick the relevant scenario to create each table, from among:
VeryHigh High  Moderate Low*/
	set seed 18
	** Set up the Parameters
	{
	if "`Correlations'"=="VeryHigh" {
		global Delta_Pk 1000
		global Delta_P 13.3
	}
	else if "`Correlations'"=="High" {
		global Delta_Pk 10000
		global Delta_P 133
	}
	else if "`Correlations'"=="Moderate" {
		global Delta_Pk 30000
		global Delta_P 200
	}
	else if "`Correlations'"=="Low" {
		global Delta_Pk 50000
		global Delta_P 1000
	}
	noisily display "`Correlations' correlations, Delta_Pk: $Delta_Pk Delta_P: $Delta_P"
	}
```

```
** Make the datasets
quietly {
	** Genderate prices and wages
	gen P = runiform(1333-$Delta_P, 1333+$Delta_P)
	gen Pk = runiform(200000-$Delta_Pk, 200000+$Delta_Pk)
	gen lP = ln(P)
	gen lPk = ln(Pk)
	global Delta_w 10000 /*Half the Delta of wages. Default = $10,000/worker-
year.*/
	gen w = runiform(16667-$Delta_w, 16667+$Delta_w) /* Default: Mean =
$50,000/worker-year.*/
	gen lnw = ln(w)
	gen WageBill = L*w
	foreach series in star end { /*Caculate V, J, r etc. for each series.*/
		gen J_`series' = Pk*K_`series'
		gen V_`series' = P*Q_`series'
		gen Profit_`series'=V_`series'-WageBill
		gen r_`series'=Profit_`series'/J_`series'
		gen a_`series' = WageBill/V_`series'
		gen JV_ratio_`series' =J_`series'/V_`series'
		gen lnV_`series'=ln(V_`series')
		gen lnJ_`series'=ln(J_`series')
		gen lnr_`series'=ln(r_`series')
		gen VMPL`series' = alpha*P*Q_`series'/L
		gen QoL`series' = Q_`series'/L
		gen VoL`series' = V_`series'/L
		gen lnTFP_factorprices`series'=a_`series'*lnw+(1-a_`series')*ln(r_`series')
	}
	format V* w P* WageBill Profit* J_* %9.0fc
	sum /*These provide the shares appearing in Table 6. */
	sum a_star
```

```
                local a_bar = r(mean)
        gen lnM_star = `a_bar'*lnw+(1-`a_bar')*lnr_star
        sum lnM_star
                local mean = r(mean)
                local sd = r(sd)
        gen z = rnormal(`mean',`sd')
        gen W_phi0 = z
        gen W_phi25 = 0.25*lnM_star + 0.75*z
        gen W_phi50 = 0.5*lnM_star + 0.5*z
        gen W_phi90 = 0.9*lnM_star + 0.1*z
    }
    ** Examine the data, expecially the correlations
    quietly {
        summarize w P Pk r* L K* J* Q* V* WageBill a* Profit* J_* JV* lnM* z
lnTFP* /*J/V is approximately 3. */
        noisily summarize lnd lnL lnK_star lnK_end lns lnQ_star lnQ_end P Pk w r_star
J_star V_star a_star lnM_star
        sum r_*, det
        count if r_star<0 /*We have some loss-making firms, but only in the low-
correlation case. This does not matter for any result presented in the paper.*/
        ** Pull out the correlations between physical and value terms (Table 3, Panel A)
        pwcorr lnK_star lnJ_star
        pwcorr lnQ_star lnV_star
        pwcorr lnd lnM_star lnTFP_factorpricesstar
}

    ** REGRESSIONS:
    display "HERE ARE THE REGRESSIONS. CORRELATION IS `Correlations'."

    ** 1. OLS, no endogeneity bias.  Table 3, Panels B-D.
    quietly {
```

```
	display "1. OLS, no endogeneity.  Table 3, Panels B-D."
	regress lnV_star lnL lnJ_star /*Panel B.*/
	regress lnV_star lnL lnJ_star lPk lP /*Panel C.*/
	regress lnV_star lnL lnJ_star lPk /*Panel D.*/
}

** 2. IV. Table 3, Panel E.
quietly {
	display "2. Instrumental Variables. Table 3, Panel E."
	ivregress 2sls lnV_star lnL (lnJ_star=lnK_star), first
}

** 2b. IV - the Griliches and Mairesse (1998) way
quietly {
	gen lnrental=lnr_end+lPk`'
	display "2b. IV - the Griliches and Mairesse way"
		ivregress 2sls lnV_star (lnL lnJ_star= lnw lnr_star), first
		ivregress 2sls lnV_star lnL (lnJ_star= lnw lnr_star), first
		ivregress 2sls lnV_end lnL (lnJ_end= lnw lnrental), first
		ivregress 2sls lnQ_end lnL (lnK_end= lnw lnrental), first
}

** 3. OLS as we approximate the accounting identity. Tables 5 (and 7).
quietly {
	pwcorr lnM_star W_phi*
	display "3. Full Accounting Identity, no endogeneity"
	quietly foreach n in 0 25 50 90 {
		regress lnV_star W_phi`n' lnJ_star lnL
			est store est`n'
	}
	quietly regress lnV_star lnM_star lnJ_star lnL
```

```
                est store estM
            quietly regress lnV_star lnJ_star lnr_star lnL lnw
                est store estID
            est table est*, keep(lnL lnJ_star W_* lnM lnw lnr_star) b(%4.3f) se(%5.4f)
stats(r2 N) /*Estimates for Table 5 (and 7).*/
            foreach est in 0 25 50 90 M ID {
                est restore est`est'
                display "Estimated returns to scale for est`est'"
                lincom lnJ_star+lnL
            }
        }
        ** Display the numbers needed to calculate the EiV and OVB biases (Tables 4 and 6).
        quietly {
            gen rho_to_Beta = -beta*lPk
            sum a_star alpha lnL rho_to_Beta /*Provides the shares needed for table 6.*/
            quietly corr lnL lnJ_star lnM_star rho_to_Beta if r_star>0, c
            matlist r(C), format(%10.8f) /*covariance matrix for the OVB and EiV
Calculations (for those with r>=0, a stipulation which only matters in the low-correlation case
because r>0 otherwise).*/
        }
        drop P - ln
}
```

Appendix E: Stata Code for Section 4.D.

```
clear all

set obs 100000

set seed 19


global SD_L 0.1

gen lnL = rnormal(ln(20),$SD_L) /*Random labor input. */
```

```
	gen L = exp(lnL)
quietly foreach Case in High Medium Low Negligible {
	if "`Case'"=="High" {
		global Range_w = 1000
		global Range_r = 0.05
	}
	else if "`Case'"=="Medium" {
		global Range_w = 500
		global Range_r = 0.02
	}
	else if "`Case'"=="Low" {
		global Range_w = 200
		global Range_r = 0.01
	}
	else {
		global Range_w = 100
		global Range_r = 0.001
	}
	global Range_a = 0.05
	gen a = runiform(0.75-$Range_a,0.75+$Range_a)
	gen w = runiform(16667-$Range_w, 16667+$Range_w) /* Default: Mean =
$16,667/worker-year.*/
	gen r = runiform(0.1-$Range_r, 0.1+$Range_r)
```

```
	gen V = (w*L/a)

	gen Profit = V - w*L

	gen lnProfit = ln(Profit)

	gen J = Profit/r

	gen lnJ = ln(J)

	gen lnV = ln(V)

	gen lnM_bar = 0.75*ln(w)+0.25*ln(r)

	display "Case is `Case'"

	sum w r a lnM_bar


	noisily regress lnV lnL lnJ

	drop a-lnM_bar

}
```