NBER WORKING PAPER SERIES

THE RELATION BETWEEN PRICE AND MARGINAL COST IN U.S. INDUSTRY

Robert E. Hall

Working Paper No. 1785

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 1986

I am grateful to Michael Knetter and David Bizer for outstanding assistance and to collegues too numerous to list, for helpful comments. The research reported here is part of the NBER's research program in Economic Fluctuations. Any opinions expressed are those of the author and not those of the National Bureau of Economic Research.

NBER Working Paper #1785
January 1986

The Relation between Price and Marginal Cost in U.S. Industry

ABSTRACT

An examination of data on labor input and the quantity of output reveals that most U.S. industries have marginal costs far below their prices. The conclusion rests on the empirical finding that cyclical variations in labor input are small compared to variations in output. In booms, firms produce substantially more output and sell it for a price that exceeds the costs of the added inputs. The paper documents the disparity between price and marginal cost, where marginal cost is estimated from variations in cost from one year to the next. It considers a wide variety of explanations of the findings that are consistent with competition, but none is found to be plausible.

Robert E. Hall
Hoover Institution
Stanford University
Stanford, California 94305

Introduction

A competitive firm equates its marginal cost to the market price of its product. The equality of marginal cost and price is a fundamental efficiency condition for the allocation of resources. When the condition holds, the purchasers of the product equate their marginal rates of substitution to the corresponding marginal rates of transformation. By contrast, under monopoly or oligopoly, the allocation of output will be inefficient because price will exceed marginal cost.

This paper derives and implements a method for estimating the ratio of price to marginal cost. The method is different from the one used in most previous investigations--instead of assuming profit maximization and estimating the slope of the demand schedule (as in Rosse (1970)), it looks at actual changes in costs. Further, the method makes no assumptions about the cost function; it is completely nonparametric. In its simplest form, it estimates the ratio of price to marginal cost directly from data on price, output, and the quantities and prices of inputs. This form proves an exact basis for a test of the hypothesis that price equals marginal cost. It also can provide the basis for estimation in a non-competitive setting when it is plausible that price is a constant markup over marginal cost, as it would be for a seller facing a demand schedule with constant elasticity.

The results of applying the estimation method to data for total

manufacturing and to 21 two-digit industries give a strong conclusion: Price far exceeds marginal cost for manufacturing as a whole and for most two-digit industries. For total manufacturing, the gap between price and marginal cost is about 63 percent of marginal labor cost. In some two-digit industries, such as paper and food and beverages, the gap is more than double marginal labor cost. The hypothesis that marginal cost is equated to price is strongly rejected for total manufacturing and for 16 out of the 21 two-digit industries.

The paper gives substantial attention to possible specification and data problems that might explain the findings without invoking a failure of the equality of marginal cost and price. First, it shows that the estimation method is robust to cyclical errors in measuring wages. As long as the average wage (or, more precisely, the average factor share of labor) is correctly measured in the long run, systematic measurement errors are essentially harmless. If labor contracts call for wage-smoothing, for example, the estimation method still works. Similarly, if the effective marginal cost of labor varies relative to the wage because of adjustment costs, the method also still works. A related argument shows that the method is essentially immune to biases from price rigidity.

Biases arising from measurement errors in output can be avoided by the use of an appropriate instrumental variable. For cyclically-sensitive industries, total real GNP works well. In industries where the bulk of output variation is idiosyncratic, it is not possible to find a good instrument.

Biases arising from measurement errors in labor input are more

of a problem. The most plausible source of such errors is unmeasured fluctuations in effort per hour of work. The estimation method is compromised only if the short-run elasticity of the supply of work effort is a significant fraction of the short-run elasticity for hours.

A related explanation of pro-cyclical productivity variations relies on labor aggregation. Suppose that a given capital stock equips a number of workers during the daytime shift, and the same capital equips a smaller number of workers paid a higher wage during the night shift. If employment during the night shift is the principal method used to vary the level of output, then productivity calculations will show a spurious procyclical element if the hours of both types of workers are added together. Night workers have a higher marginal product than do day workers. However, the numerical magnitude of this bias is far too small to account for the findings of the paper. Another potential explanation for the finding of marginal cost below price is increasing returns to scale. In a sense, this explanation is complementary to the basic conclusion of the paper, since a firm with increasing returns that equated its marginal cost to price would operate at a loss. Increasing returns virtually requires a market structure such that marginal cost falls short of price. In any case, a modification of the basic equation of the paper shows that increasing returns is only strongly evident in a few industries, such as electricity generation. In most industries, constant returns is supported by the data.

1. The method

Consider a firm that produces output Q with capital K and labor N. Assume constant returns to scale in K and N. Then the production function can be written in intensive form as:

$$(1.1) \qquad Q/K = e^{\theta t} f(N/K)$$

The intensive production function, $f(\ )$, is concave; θ is the rate of Hicks-neutral technical progress or rate of growth of total factor productivity. Let q be the log of the output/capital ratio or capacity utilization rate ($q = \log(Q/K)$) and let n be the log of the labor/capital ratio ($n = \log(N/K)$). Taking the time derivative and approximating with discrete changes gives

$$(1.2) \qquad \Delta q = \theta + \frac{N e^{\theta t} f'(N/K)}{Q} \Delta n$$

Marginal cost is the ratio of the wage to the marginal product of labor:

$$(1.3) \qquad \text{Marginal cost} = \frac{w}{e^{\theta t} f'(N/K)}$$

The hypothesis that price is a constant ratio to marginal cost can be expressed as

$$(1.4) \qquad p = \frac{1}{1-\beta} \frac{w}{e^{\theta t} f'(N/K)}$$

The parametrization of the ratio as $1/(1-\beta)$ is chosen for convenience. If β is 0.5, for example, price is double marginal cost. Putting the hypothesis that price is proportional to marginal cost into the expression for the rate of growth of output gives

$$(1.5) \qquad \Delta q = \theta + \frac{1}{1-\beta} \frac{wN}{pQ} \Delta n$$

The fraction wN/pQ is just labor's share in total revenue; I will call it α. Making this substitution and multiplying the equation by $1-\beta$ gives

$$(1.6) \qquad \Delta q - \alpha \Delta n = (1-\beta)\theta + \beta \Delta q + u$$

I have added a random disturbance, u, to take account of the facts that price is not literally proportional to marginal cost and technology does not always advance at exactly the same rate, θ.

Equation 1.6 is the basic idea of the paper. It can be explained in the following way. When price equals marginal cost, the revenue share of labor, α, measures the elasticity of output with respect to labor, independent of the form of the technology. Subtracting α times the log change in labor input from the observed log change in output would yield just the rate of technical progress, under

marginal cost pricing. However, if the revenue share of labor understates the elasticity of output with respect to labor, because price exceeds marginal cost, then the left-hand side will contain a component related to the change in output. In fact, the coefficient of the log-change in output is precisely the parameter, β, that controls the departure of price from marginal cost. The observation that, when price and marginal cost are equal, the left-hand side of equation 1.6 measures the growth of producitivty was first made in a famous paper by Robert Solow (1957). However, the applicaton of the formula when price is different from marginal cost is new, so far as I know.

Solow's method, when applied to intensive data, does **not** rest on any assumption that the firm is using its equilibrium amount of capital. Rather, it only uses an estimate of the elasticity of output with respect to labor input, α, and that estimate can be made accurately from labor's share of revenue, provided only that the firm is equating the marginal product of labor to the product wage. There is no implicit assumption that the marginal product of capital is being equated to the real rental price of capital. Similarly, the technique proposed here requires only that the firm sets marginal cost to a proportion of price, given whatever is its capital stock, and makes no assumption about how the firm chooses its capital stock.

Estimation of the basic equation will require the hypothesis that the disturbance, u, is uncorrelated with changes in output, or, in the case of instrumental variable estimation, that u is uncorrelated with the instrument. Since output is a highly cyclical variable, and the instrument to be used is a cyclical aggregate demand variable, a

vigorous defense of this hypothesis is required. That is, my method depends critically on the hypothesis that there is no true cyclical variation in the Hicks-neutral rate of productivity growth. There is no denying that producitivity is pro-cyclical in the sense that output per employee-hour is high when output is high. Nor, for that matter, is it in dispute that the year-by-year rate of total factor producitivty growth calculated by Solow's method is equally cyclically variable. However, both of these well-known propositions are perfectly consistent with the basic hypotheses I wish to maintain. It is eminently likely that firms hoard labor during contractions, which is the usual explanation for cyclical variations in output per employee-hour. However, labor hoarding should not cause cyclical fluctuations in total factor productivity; one of the many virtues of Solow's technique is its robustness in the presence of labor hoarding.

Consider a cyclical contraction where output falls substantially but labor falls by much less because of hoarding. In such a contaction, marginal cost falls to a low level. The cost of incremental output is low because it can be produced simply by putting hoarded workers back to work, with little increase in payroll costs. If price equals marginal cost, the revenue share of labor rises dramatically because the price falls to the low level of marginal cost. Consequently, no special cyclical effect appears in equation 1.6 because the rise in α offsets the low value of Δn. Of course, the revenue share of labor does not rise very much, if at all, in contractions, so the Solow calculation gives a substantial decline in productivity in each contraction. But my point is that the reason for this finding is the failure of marginal cost pricing. The

pervasive belief in cyclical fluctuations in productivity is really a pervasive belief that price does not track marginal cost. My equation 1.6 is the most plausible explanation for cyclical fluctuations in productivity, in this interpretation.

I would also offer the positive argument in favor of my basic identifying hypothesis that the process of technical change should logically proceed fairly smoothly. The index of productivity is a feature of the production function. Contractions are not periods when firms forget their best prodution techniques and retreat to less efficient ones. Rather, the process of creating and installing new techniques should proceed at very much the same pace through cyclical expansions and contractions.

Under the basic identifying hypothesis that true shifts in productivity are unrelated to cyclical fluctuations in real GNP, estimation of the degree of departure from marginal cost pricing is a simple matter of comparing the cyclical behavior of the Solow residual, $\Delta q - \alpha\Delta n$, to the behavior of the rate of growth of output. To the extent that periods of rising output are ones when the actual growth of output exceeds the amount expected from observations on the revenue share, α, applied to labor growth, Δn, price is shown to exceed marginal cost.

2. Value added

In addition to the labor and capital considered in the previous section, firms use materials and other intermediate products as

inputs to production. Were time series data on other inputs available, it would be a simple matter to add additional terms to equation 1.6, each containing a factor share multiplying a rate of growth of an input. However, full input-output data are not available on an annual basis for U.S industries. Rather, research of this type must make use of annual data on nominal and real value added. This section modifies the earlier analysis to deal with that problem.

In this section, variables with *s signify measures of the theoretical ideal: Q^* is true gross output, q^* is the log of the ratio of Q^* to capital, p^* is the actual price of output, μ^* and α^* are the factor shares of materials and labor relative to the value of gross output, p^*Q^*, θ^* is the rate of Hicks-neutral technical progress in the production function relating gross output to all inputs, and β^* is the parameter governing the ratio of the actual price to marginal cost. Also, z is the price of materials, M is the quantity of materials employed and m is the log of the materials-capital ratio. Then a simple extension of equation 1.6 shows how β^* could be measured in this setup:

$$(2.1) \qquad \Delta q^* - \alpha^* \Delta n - \mu^* \Delta m = \theta^*(1-\beta^*) + \beta^* \Delta q^*$$

However, the output measure that is available is not Q^*, gross output, but is Q, real value added. The rate of growth of the ratio of real value added to the capital stock is

$$(2.2) \qquad \Delta q = \frac{\Delta(Q/K)}{Q/K} = \frac{p^*\Delta(Q^*/K) - z\Delta(M/K)}{p^*Q^*/K - zM/K}$$

$$= \frac{\frac{\Delta(Q^*/K)}{Q^*/K} - \frac{zM}{p^*Q^*}\frac{\Delta(M/K)}{M/K}}{1 - \frac{zM}{p^*Q^*}}$$

$$= \frac{\Delta q^* - \mu^*\Delta m}{1 - \mu^*}$$

This relation can be used to eliminate the unobserved Δq^* from equation 2.1:

$$(2.3) \qquad \Delta q - \alpha\Delta n = \theta(1-\beta^*) + \beta^*\Delta q + \beta^* \frac{\mu^*}{1-\mu^*} \Delta m$$

Here α is the labor's share in value added and θ is the rate of technical progress stated in labor-capital augmenting form ($\theta = \theta^*/(1-\mu^*)$). Equation 2.3 says the following: If the growth of the materials-capital ratio, Δm, is uncorrelated with the growth of the output-capital ratio, Δq, then the regression of the Solow residual on the rate of growth of the output-capital ratio, Δq, will reveal the price-marginal cost parameter, β^*. However, to the extent that Δm is positively correlated with Δq (the likely case), then the regression coefficient will overstate β^*. In particular, in the case where the ratio of materials to output is technologically fixed, then $\Delta m = \Delta q$, and equation 2.3 becomes

$$(2.4) \qquad \Delta q - \alpha\Delta n = \theta(1-\beta^*) + \beta\Delta q$$

where $\beta = \beta^*/(1 - \mu^*)$. In this case, the estimated coefficient, β, has the interpretation of the ratio of the gap between price and marginal cost to value added per unit of output. β exceeds β^* by a factor related to the importance of materials inputs. Given an estimate of β, the corresponding estimate of the ratio of price to marginal cost can be recovered by multiplying by 1 minus the factor share of materials.

Estimates of β are interesting in their own right, without adjustment for the share of materials. In the first place, they measure the price distortion relative to value added. Second, they provide the best guide to the overall degree of excess of price over marginal cost for the economy as a whole. Suppose every industry had the same β and the same μ^*. Then the ratio of the price of a particular final good to the total marginal cost of production, counting all stages, would be β. The β^* for any given industry would understate the distortion of any given price because it would not count the distortion built into materials prices.

The discussion in this section made the implicit assumption that the change in real value added was computed each year using the previous year's prices as the base prices (see equation 2.2). In effect, it assumed the use of a Divisia index of real value added. In the U.S. national income accounts, base prices are changed about once a decade. I know of no reason to think that the low frequency of base changes has any important influence on the results obtained by the technique in this paper.

3. *Data*

I have obtained results for total manufacturing and for 21 two-digit industries. The data are:

- Q: Real value added, U.S. NIPA
- K: Net real capital stock, BEA.
- p: Implicit deflator with indirect business taxes removed (Ratio of nominal value added less IBT to real value added)
- N: Hours of work of all employees, U.S. NIPA
- w: Total compensation divided by N

Note that the data are chosen to eliminate tax wedges as a source of departures of marginal cost from price. The price level is measured net of sales and other taxes, and the wage is measured gross of social security, fringes, and other costs incurred by the employer.

The NIPA do not report hours of all employees by industry after 1978 or before 1948. Hence, the period studied is 1949 through 1978. In addition, certain industries underwent definitional changes in 1973. For those industries, I omitted the 1972-73 change from the estimation process. The industries are: Lumber (SIC 24), Furniture (25), Chemicals (28), Rubber (30), Primary Metals (33), Fabricated Metals (34), Electrical Machinery (36), and Instruments (38).

4. *Results*

Total manufacturing

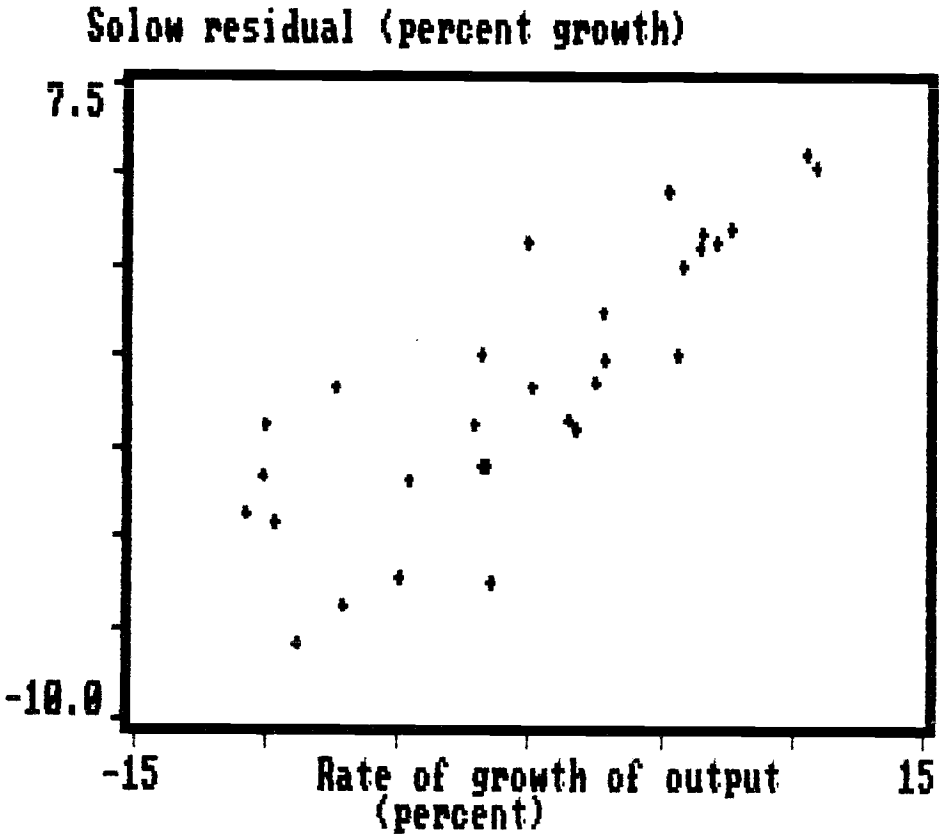
Table 1 shows the construction of the dependent variable and Figure 1 shows a scatter diagram of the Solow residual against the change in output. A strong positive relation is immediately apparent. There is no mistaking the fact that output grows by more than can be explained by applying the product wage as an estimate of the marginal product of labor to the observed growth in labor. Output consistently grows by more in expansions and falls by more in contractions. The most obvious explanation is that the product wage understates the marginal product of labor; that is, price exceeds marginal cost.

The slope of the relation between the Solow residual and the rate of growth of output is β, the parameter that governs the relation between price and marginal cost. Because the right-hand variable also appears in the construction of the left-hand variable, it is essential to use an estimation method that is immune to measurement errors in output. Regression estimates of β would be biased upward by purely random errors in the growth of output. My estimates are based on an instrumental variables procedure, with the rate of growth of real GNP as the instrument.

Table 1. Data for total manufacturing
(percent change or percent)

Year	Output growth	Hours growth	Labor share	Solow residual
1949	-7.4	-9.2	71.5	-0.9
1950	10.8	8.1	69.8	5.1
1951	5.6	7.9	69.7	0.1
1952	-1.7	1.9	72.7	-3.1
1953	2.5	4.4	74.0	-0.7
1954	-10.9	-8.7	74.4	-4.4
1955	7.1	5.5	71.8	3.1
1956	-5.1	1.5	73.9	-6.2
1957	-4.7	-1.7	74.6	-3.4
1958	-10.2	-9.2	76.1	-3.2
1959	10.4	6.6	73.5	5.6
1960	-2.1	-0.4	75.4	-1.8
1961	-1.9	-2.6	75.3	0.1
1962	5.8	4.4	74.4	2.5
1963	5.2	0.9	73.1	4.5
1964	2.7	2.1	72.7	1.2
1965	1.7	5.2	70.8	-2.0
1966	-1.5	6.7	71.5	-6.3
1967	-7.1	-0.3	73.0	-6.9
1968	.0	1.3	73.2	-0.9
1969	-1.9	1.6	75.4	-3.1
1970	-9.8	-6.6	77.9	-4.6
1971	-0.1	-4.2	75.6	3.1
1972	6.5	4.1	75.1	3.4
1973	7.6	5.5	76.0	3.5
1974	-8.9	-1.3	78.8	-7.9
1975	-10.1	-10.9	75.5	-1.9
1976	6.4	4.6	74.3	3.0
1977	2.8	3.9	73.8	-0.1
1978	1.4	4.3	74.4	-1.8

Figure 1. The Solow residual and the rate of growth of output, total manufacturing, 1949-1978



For the whole period, 1949 through 1978, the estimated slope is

$$(4.1) \qquad \Delta q - \alpha \Delta n = -.0071 + .385 \Delta q \qquad (.070)$$

Standard error: 2.2 % Durbin-Watson statistic: 1.40

The hypothesis that price and marginal cost are equal (β=0) is overwhelmingly rejected. The implied ratio of the deflator to marginal labor cost, $1/(1-\beta)$, is 1.63. The gap between price and marginal cost is 63 percent of value added. For the manufacturing industry as a unit, value added is 56.9 percent of sales, netting out sales within manufacturing. Thus β^* is no less than 0.385 x .569 = .219 and price is at least 1/(1-.219) = 1.28 times total marginal cost. However, this calculation understates the total price distortion for manufactured goods because it omits distortions in the prices paid for materials and produced inputs and because it assumes that the change in materials input is perfectly correlated with the change in output..

Two-digit industries

Results for selected two-digit industries are presented in Table 2.

Table 2. Estimates for two-digit industries, 1949-78

SIC code	Description	Slope (IV estimates)	Std. err. (IV estimates)	DW (IV estimates)	Slope (OLS)	Value added share (%)
20	Food and beverages	0.683 (.198)	1.50	1.53	0.837 (.067)	29.7
21	Tobacco	0.158 (.718)	5.61	2.29	0.635 (.032)	44.2
22	Textiles	0.059 (.245)	5.20	1.89	0.556 (.084)	43.5
23	Apparel	0.173 (.179)	3.28	1.99	0.275 (.095)	46.9
24	Lumber	0.152 (.204)	5.18	1.80	0.591 (.079)	54.7
25	Furniture	0.271 (.085)	2.85	2.22		41.9
26	Paper	0.661 (.044)	1.51	1.36		48.4
27	Printing and publishing	0.383 (.233)	1.74	1.73	0.584 (.069)	51.5
28	Chemicals	0.831 (.052)	1.45	1.74		46.2
29	Petroleum refining	0.639 (.159)	3.31	1.10	0.874 (.077)	10.5
30	Rubber	0.289 (.092)	3.75	2.41		43.7
31	Leather	0.424 (.159)	4.52	2.66	0.405 (.081)	49.8
32	Stone, clay, and glass	0.536 (.048)	1.49	2.02		51.6
33	Primary metals	0.512 (.033)	1.91	2.36		43.9
34	Fabricated metals	0.291 (.066)	2.40	1.43		45.4
35	Machinery (non-elec.)	0.273 (.056)	2.14	2.23		50.9
36	Machinery (electrical)	0.301 (.081)	3.23	2.35		50.6
38	Instruments	0.251 (.081)	2.85	2.52		56.0
39	Miscellaneous	0.485 (.098)	2.39	2.50		44.5
48	Communication	0.709 (.254)	1.91	1.27	0.529 (.110)	76.2
49	Elec., gas, and sanitary	0.874 (.319)	1.01	0.48	0.965 (.066)	50.3

Notes: The slope is the coefficient of the rate of change of output. The dependent variable is the Solow productivity residual. OLS estimates are given in those cases where the standard error of the estimated slope is greater than 0.150.

In all industries save textiles, tobacco, apparel, lumber, and printing and publishing, the hypothesis of price equal to marginal cost is soundly rejected. In eight of the 21, β exceeds 0.5, which means that the price distortion exceeds marginal labor cost itself.

In most of the industries where imprecise values of β are found, the problem appears to be that real GNP is a poor instrument; most demand shifts are idiosyncratic to the industry and are not correlated with overall economic activity. It is interesting to look at ordinary least squares results for these industries, even though they may be biased because of errors in measuring output. The OLS results are given on the right of Table 2. In all five of the industries where the hypothesis of equality of marginal cost and price could not be rejected on the basis of the instrumental estimates, the hypothesis is rejected on the strength of the OLS results. Moreover, in all but one industry, apparel, the OLS estimate of β exceeds 0.5.

5. *Analysis of specification errors*

A number of explanations of cyclical fluctuations in productivity come to mind that would represent specification errors in terms of the theory used in this work. First, employers may pay their workers under a wage-smoothing arrangement. Under such an arrangement, the wage equals the long-run opportunity cost of time, but does not track short-run fluctuations in labor-market conditions. Second, employers may incur adjustment costs, in which case the wage does not measure all components of marginal cost related to labor. Third, prices may not be fully flexible, even though they do not differ from marginal cost on the average. Fourth, labor input may be measured incorrectly. Fifth, hours of work may not be aggregated correctly. Sixth, the technology may have increasing returns to scale.

The first part of this section develops an argument that none of the first three specification errors could explain the basic finding of the paper. The first two errors make the measured wage differ from the true effective wage over the cycle, but not in the long run. Such errors bias the estimate of β only in a certain second-order way, and, in any case, the bias is downward. Neither could explain the finding of strongly positive values of β. Similarly, price rigidity creates only a tiny bias in the estimate and could not explain a strong positive β.

The general intuition behind this conclusion can be explained easily. Consider a situation where marginal cost is equated to price and all the other assumptions necessary to justify the method of the

paper hold. Then

$$(5.1) \qquad r = \Delta q - \alpha \Delta n = \theta$$

and the regression of r on Δq will yield a coefficient of zero. Now suppose that an erroneous measure, $\hat{\alpha}$, is used in place of α in the computation of the residual. It may differ from α because of wage smoothing, adjustment costs, price rigidity, or any other reason. Then the measured residual, r, will be

$$(5.2) \qquad \begin{aligned} \hat{r} &= \Delta q - \hat{\alpha} \Delta n \\ &= \frac{\alpha - \hat{\alpha}}{\alpha}(\Delta q - \theta) + \theta \end{aligned}$$

Then the regression of $\hat{r}$ on Δq will yield the coefficient,

$$(5.3) \qquad \hat{\beta} = \frac{\mathrm{Cov}(\frac{\alpha - \hat{\alpha}}{\alpha}(\Delta q - \theta), \Delta q)}{\mathrm{V}(\Delta q)}$$

The covariance in the numerator has two terms. The first,

$$(5.4) \qquad \mathrm{Cov}(\frac{\alpha - \hat{\alpha}}{\alpha}\Delta q, \Delta q)$$

will be zero if $(\alpha - \hat{\alpha})/\alpha$ and Δq both have mean zero and $(\alpha - \hat{\alpha})/\alpha$ and $(\Delta q)^2$ are uncorrelated. The lack of correlation is likely to hold under rather general conditions. Unbiasedness of $\hat{\alpha}$, in the sense of zero mean of $(\alpha - \hat{\alpha})/\alpha$, is the substantive requirement.

The second term in the covariance,

(5.5) $$-\theta \mathrm{Cov}\left(\frac{\hat{\alpha}-\alpha}{\alpha}, \Delta q\right)$$

is unlikely to be zero, assuming that the measurement error, $\hat{\alpha}-\alpha$, tracks movements in output. However, the term is likely to be small because it is multiplied by the rate of productivity growth, θ. Moreover, if the covariance of the departure of the share, $(\hat{\alpha}-\alpha)/\alpha$, with the growth of output, Δq, is positive, then the bias will be downward. In other words, if the share tends to be understated when output is growing, then $\hat{\beta}$ will be biased downward.

I summarize these conclusions in a

Theorem

If

(i) The true residual, $\Delta q - \alpha \Delta n$, is equal to the constant rate of productivity growth, θ, and

(ii) The measured share, $\hat{\alpha}$, is unbiased in the sense that $E[(\hat{\alpha}-\alpha)/\alpha] = 0$, and

(iii) q obeys a stationary stochastic process with constant mean and a symmetric distribution about the mean, and

(iv) $(\hat{\alpha}-\alpha)/\alpha$ depends linearly on past, present, and future values of q,

then the regression coefficient for the measured residual on the rate of change of output is

$$\hat{\beta} = -\theta \frac{\mathrm{Cov}\left(\frac{\alpha-\hat{\alpha}}{\alpha}, \Delta q\right)}{V(\Delta q)} \tag{5.6}$$

The proof is sketched above. Note that symmetry of the distribution of q about its mean is sufficient to eliminate the covariance of $(\alpha-\hat{\alpha})/\alpha$ and $(\Delta q)^2$, since this covariance will depend only on third moments of the distribution of q, which are zero by virtue of symmetry.

The errors arising from the first two sources considered here arise from the wage. When the other components of the share (price, quantity, and employment) are measured accurately, then the proportional error in the share is equal to the proportional error in the wage. If w is the true wage (in the sense of the effective marginal cost of labor) and $\hat{w}$ is an erroenous measure of it, then

$$\frac{\alpha-\hat{\alpha}}{\alpha} = \frac{w-\hat{w}}{w} \tag{5.7}$$

Similary, when only the price is measured with error, the proportional error in the share is the proportional error in the price, with its sign reversed.

Wage smoothing

Martin Neil Baily's (1974) pioneering paper pointed out the advantage to workers of earning smoothed wages. When workers

cannot use credit markets as easily as employers can, then it makes sense to decouple earnings from labor-market fluctuations. In the extreme version, workers receive a predetermined real annual income, unrelated to the amount of work they do and unrelated to the value of their time. Though such an arrangement could be examined with the aid of the theorem, I have taken a less extreme view. Suppose that workers receive a guaranteed hourly wage, but the wage is paid only for hours actually worked. Those hours are determined not by equating the value of the marginal product of labor to the wage, but rather by equating the value of the marginal product to the marginal value of time. The error in measuring the share arises because the effective wage is the value of time, but the measured wage is the predetermined contract wage.

Suppose that the wage error can be written as

$$(5.8) \qquad \frac{w-\hat{w}}{w} = \delta(n - \bar{n})$$

where δ is the reciprocal of the elasticity of labor supply; that is, the elasticity of the marginal value of time with respect to the amount of work. The elasticity δ is a short-run concept, so it is reasonable to assume that the substitution effect dominates the income effect and δ is not too large. The percent error in the measured share is the same, and can be written in terms of output:

$$(5.9) \qquad \frac{\alpha-\hat{\alpha}}{\alpha} = \delta(n - \bar{n}) = \frac{\delta}{\alpha}(q - \bar{q})$$

Application of the theorem will make use of

$$(5.10) \qquad \frac{Cov(q,\Delta q)}{V(\Delta q)} = \frac{1}{2}$$

which is true for any stationary times series. Then the theorem implies that the coefficient of the regression of the Solow residual on the rate of change of the output-capital ratio is:

$$(5.11) \qquad \hat{\beta} = -\frac{\delta\theta}{2\alpha}$$

The first thing to say is that this is a small number, compared to the estimates of β presented earlier in this paper. If δ is 1, θ is .03, and α is 0.7, then $\hat{\beta}$ is - 0.02. Second, $\hat{\beta}$ is negative. The bias from wage-smoothing is small and negative; it can form no part of the explanation of the finding of positive values of β around 0.5.

Let me stress again the basic reason for this finding. As long as the Solow residual is computed with an ***unbiased*** estimate of α, the bias in $\hat{\beta}$ is necessarily small. To get a substantial positive value of $\hat{\beta}$, it would be necessary to use an estimate of α that was systematically too small. Marginal cost chronically below price is the most obvious source of a downward bias in α.

Costly adjustment

A second potential source of departure of the effective marginal cost of labor from the quoted wage could arise from adjustment costs

for labor. Mark Bils (1985) has investigated the ways that adjustment costs affect marginal cost.

In a simple setup where costs of adjustment are quadratic in the proportional change in effective labor input, the marginal cost of adding an hour of labor input in year t is:

$$(5.12) \qquad w_t = \hat{w}_t(1 + \gamma\Delta\tilde{n}_t - \gamma\Delta\tilde{n}_{t+1})$$

Here $\hat{w}$ is the measured hourly wage and $\tilde{n}$ is effective labor input, that is, the log of hours of work measured in efficiency units:

$$(5.13) \qquad \tilde{n}_t = \frac{\theta}{\alpha}t + n_t$$

The paremeter γ has the following interpretation: If γ is 1, at a time when employment has risen by 10 percent, the marginal adjustment cost of labor is 1/10 of the direct wage cost.

From the adjustment cost model, it is easy to derive that:

$$(5.14) \qquad \frac{\alpha-\hat{\alpha}}{\hat{\alpha}} = \frac{\gamma}{\alpha}(\Delta q_t - \Delta q_{t+1})$$

The crucial covariance for the application of the theorem is

$$(5.15) \qquad \mathrm{Cov}(\Delta q_t - \Delta q_{t+1}, \Delta q_t) = (1-\rho)V(\Delta q)$$

where ρ is the serial correlation of Δq. Then the theorem says that the coefficient for the regression of the Solow residual on the rate of change of output is:

(5.16) $$\hat{\beta} = -\frac{\theta\gamma}{\alpha}(1-\rho)$$

Again, the coefficient is small and negative. If θ is 0.03, γ is 2, ρ is -.3, and α is 0.7, then $\hat{\beta}$ is -0.11.

As with wage smoothing, adjustment costs do not cause the measured share of labor to depart from the true share (computed shadow cost of hours of work) in the long run. Hence the bias from the error in measuring the share is small.

Price rigidity

A good deal of thought and evidence points in the direction of price rigidity. Product prices fluctuate less as demand changes than is predicted by the competitive model. Will the method of this paper suffer from an important bias if the price is rigid? The answer is no. If price rigidity is itself unbiased--if the price spends as much time below marginal cost as above marginal cost--then the method will yield an estimate of β that is essentially zero. Consider first the relation between price and output predicted by the competitive model, which can be approximated closely as

(5.17) $$p = \alpha\phi we^{-\theta t}[1 + \frac{1-\alpha}{\alpha}(q - \bar{q})]$$

Here ϕ is a constant that depends on the steady-state output-capital ratio, which depends in turn on the rental price of capital. Because the equation makes price conditional on the capital stock, the rental price does not appear explicitly. A reasonable characterization of an unbiased but rigid price simply drops the $q - \bar{q}$ term from equation 5.17:

$$\hat{p} = \alpha\phi we^{-\theta t} \tag{5.18}$$

Because many products have product wages (ratios of the wage to the implicit deflator for the product) that follow smooth trends, this type of "price equation" does well in explaining the data. Theories of price rigidity generally assume that the firm is a quantity-taker and is typically off its short-run supply schedule. For a given level of output, p from equation 5.17 is the proper price to use in productivity calculations, whereas $\hat{p}$ from equation 5.18 is what is actaully used. The error in the share is

$$\frac{\alpha - \hat{\alpha}}{\alpha} = \frac{\hat{p} - p}{p} = \frac{1 - \alpha}{\alpha}(q - \bar{q}) \tag{5.19}$$

Because the share is right on the average, the Theorem applies, and the estimated value of β is

$$\hat{\beta} = \frac{1-\alpha}{\alpha}\frac{\theta}{2} \tag{5.20}$$

Though positive, this number is invariably small. For example, if α is 0.7 and θ is 0.02, then $\hat{\beta}$ is only 0.004. Once again, a specification error that does not bias the labor share has almost no

impact on the estimate of β. The finding of strong positive values of β must come from other sources.

Problems in measuring labor input

I turn now to specification and data errors that influence labor input, n, rather than the labor share, α. In the first place, purely random errors in Δn, uncorrelated with the right-hand variable Δq, do not bias the estimate of β. However, the hypotheses that spring to mind about errors in Δn suggest they would be negatively correlated with Δq. Suppose, for example, that some workers always report 40 hours of work per week even though they work more hours when demand is strong and fewer when it is weak. Then the correlation is clearly negative. Such a negative correlation could explain the finding of positive β, since a negative correlation between Δq and errors in Δn brings positive correlation between Δq and the measured left-hand variable. In formal terms, if $\hat{\Delta n}$ is an erroneous measure of Δn, such that a fraction ψ of movements in Δn are omitted from $\hat{\Delta n}$, then the Solow residual becomes

$$\Delta q - \alpha\hat{\Delta n} = (1-\psi)\theta + \psi\Delta q \tag{5.21}$$

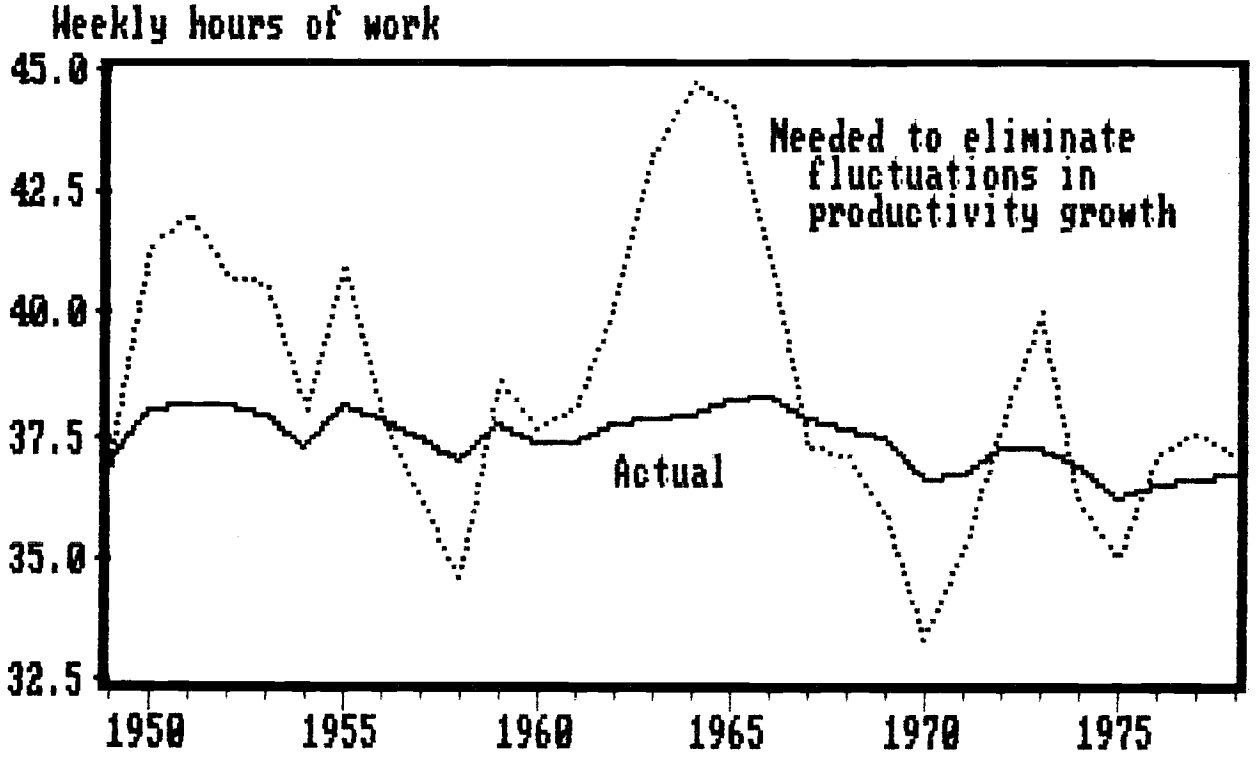
Plainly, the estimate of β is just ψ. A perfectly competitive industry would be diagnosed as having price in excess of marginal cost, when in fact the problem was the understatement of fluctuations in labor input. The likely source of errors in measuring total employee-hours is presumably in hours per worker, rather than in the count of workers.

However, changes in the number of workers account for a significant proportion of total variations in employee-hours. It turns out that the magnitude of the fluctuations in actual hours necessary to explain the finding of large β is implausibly high. Figure 2 illustrates the point for total manufacturing. The solid line shows the modest fluctuations in weekly hours per worker as measured in the NIPA. The broken line shows the huge fluctuations in hours needed to rationalize the finding of $\hat{\beta}$=0.38.

A more subtle problem of measurement of labor input would arise if labor had two dimensions, hours and effort. Suppose, for concreteness, that hours, h, and effort, f, multiply to form labor input , n. If fluctuations in effort are ignored in computing Δn, then the situation will be the same as just described for errors in measuring hours. Figure 3 is similar to Figure 2 in computing the magnitude of the fluctuations in work effort needed to explain the measured fluctuations of productivity in a competitive setting. Note in particular that effort was more than 10 percent above normal for three successive years in the mid-1960s.

The variable f can be interpreted as accomplishments per hour; then the assumption that output depends on n says that the unit of labor input is the accomplishment. Under these conditions together with competition, workers would be paid a piecerate per accomplishment equal to the marginal value of an accomplishment. One of the ways of appraising the competitive explanation of the finding of positive vlaues of β by way of unobserved variations in work effort is to ask about its implications for

Figure 2. Actual hours and hours needed to explain fluctuations in productivity, manufcturing, 1949-78



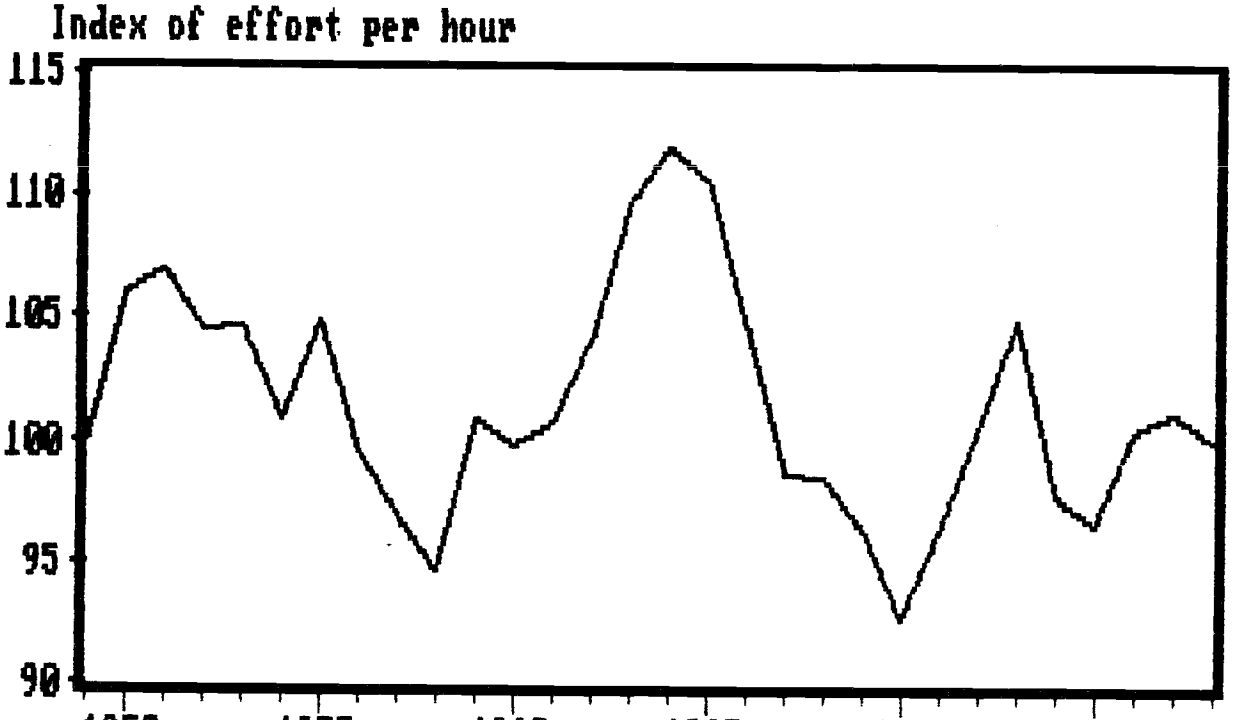


Figure 3. Work effort needed to explain fluctuations in productivity

labor supply. With the piecework technology, firms are indifferent between various combinations of hours and effort that yield the same volume of accomplishments. The split between hours and effort is based purely on the preferences of workers.

In this setting, the parameter β is interpreted as the ratio of the elasticity of the supply of effort with respect to the piecerate wage to the elasticity of the supply of hours with respect to that wage. A β of 0.5, for example, means that a decline in the piecerate brings equal percentage declines in effort and hours. In effect, workers with those preferences respond to lower piecerates by working less intensively. They could reduce their hours twice as much by continuing the same level of effort, but choose more leisure on the job and less leisure off the job.

Figure 4 shows that any competitive explanation of the finding of positive β based on errors in measuring hours or effort must rely heavily on the theory of wage smoothing. It shows the actual hourly wage and the hourly wage computed as the ratio of compensation to the adjusted measure of labor input underlying Figures 2 and 3. The inferred wage is hourly compensation per actual hour or per accomplishment. In the expansion of the mid-1960s, the inferred wage actually declined; in the highly inflationary expansion of the early 1970s, it remained level. It is highly unlikely that the market-clearing wage moved along the inferred path. Rather, the competitive explanation must assert that compensation is decoupled from hours of work or work effort. The bulges of extra, unmeasured hours in Figure 2 or the bulges of intense effort in Figure 3 were not paid for on a current basis by employers. Instead, workers provided the extra labor input in accord with of long-term agreements, if the competitive story is to be believed.

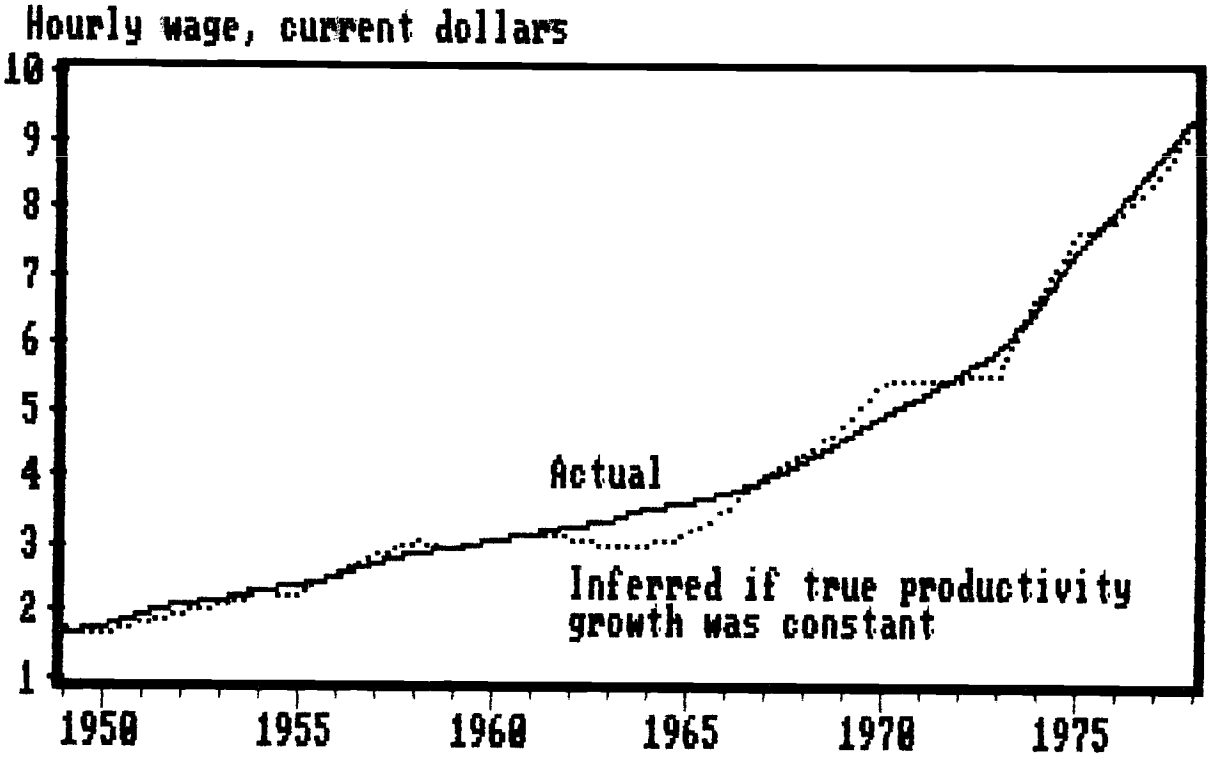


Figure 4. Actual and inferred hourly wages

Labor aggregation

One of the few attempts to explain the phenomenon of pro-cyclical productivity in terms of competitive theory was introduced by Robert Lucas (1970) and pursued by Thomas Sargent and Neil Wallace (1974). In essence, they argue as follows: The length of the work day is determined by economic considerations. Hours of work are more expensive at night, but a night shift adds to the working hours of capital. Moreover, as demand fluctuates, much of the adjustment in output may occur in the form of variations in the use of night shifts. Because an hour of work at night is more expensive, the nighttime labor/capital ratio is lower and the marginal product of labor is higher. Hence, a productivity calculation that uses the average productivity of labor to adjust for variations in labor input will understate the actual marginal product and will create the erroneous impression of procyclical productivity.

Let n_d be the labor/capital ratio for the day shift and n_g be the ratio for the night or graveyard shift. Then, under constant returns and a constant rate of growth of productivity,

$$\Delta q - \alpha_d \Delta n_d - \alpha_g \Delta n_g = \theta \tag{5.22}$$

where α_d and α_g are the elasticities with respect to the two types of labor; these could be measured as factor shares under competition. Suppose that the Solow residual is computed in terms of aggregate hours:

$$\hat{r} = \Delta q - (\alpha_d + \alpha_g)\, \Delta n \tag{5.23}$$

Two features of the night shift relative to the day shift determine the bias in measuring β:

λ: The excess sensitivity of night work relative to day work as output varies ($\Delta n_g = (1+\lambda)\Delta n_d$)

δ: The wage differential for night work ($w_d = (1-\delta)w_g$)

Using equation 5.22 to eliminate Δn from equation 5.23, and substituting the two assumptions about Δn_g and w_d just stated, I get:

$$(5.24) \qquad \hat{\beta} = \frac{\lambda\delta\alpha_d\alpha_g}{[\alpha_d+(1+\lambda)\alpha_g][\alpha_d+(1-\delta)\alpha_g]}$$

If λ is 4, δ is 10 percent, α_d is 0.5, and α_g is 0.2, then $\hat{\beta} = 0.034$. The bias from this source is trivial.

Increasing returns to scale

The general production function, possibly with increasing returns to scale, is

$$(5.25) \qquad Q = e^{\theta t}F(N,K)$$

Then

(5.26) $$\Delta q = \theta + \alpha \Delta n + (\alpha + \gamma - 1)\, \Delta k$$

where γ is the elasticity of output with respect to capital and Δk is the percent increase in K. Under constant returns, when $\alpha + \gamma = 1$, this expression boils down to Solow's equation. Now if an erroneously low estimate of α is used, $\hat{\alpha} = (1-\beta)\alpha$, say because it is measured from the labor share in the presence of marginal cost below price, then the Solow residual becomes:

(5.27) $$\hat{r} = \Delta q - \hat{\alpha}\Delta n = (1-\beta)\theta + \beta \Delta q + (1-\beta)(\alpha + \gamma - 1)\Delta k$$

A test of the inclusion of Δk in the equation can distinguish constant from increasing returns to scale, independent of the technology. Furthermore, in the presence of increasing returns, the coefficient of Δk will serve as an estimate of $1 - \beta$ times the degree of increasing returns. If the technology is Cobb-Douglas, the coefficient will have exactly this interpretation; if factor shares are reasonably constant over time, the coefficient will serve as a good average measure.

Adding the rate of growth of capital to the equations estimated in Table 2 produced the following results: The standard errors of the estimates of $(1-\beta)(\alpha+\gamma-1)$ were generally in the range of 0.1 to 0.2. In 17 out of the 21 two-digit industries, the hypothesis of constant returns was accepted at the 95 percent level. Electric-gas-sanitary (SIC 49) and leather (SIC 31) had significantly increasing returns, with coefficients of 0.377 and 0.328 respectively. Non-electrical machinery (SIC 35) and communications

(SIC 48) had significant decreasing returns. Even in those industries with evidence of increasing returns, the estimates of β itself were scarcely affected by adding Δk to the equations.

I conclude that there is little reason to believe that increasing returns to scale explain the findings of strongly positive values of β. Even if one of the equations had a value of β of close to zero together with strong increasing returns, the findings would be paradoxical, for they would imply that the firm or industry was operating at a loss. No sensible model of a private industry could explain that combination of findings.

6. *Interpretation and conclusions*

The basic fact found in this paper is neither new nor surprising. When output rises, firms sell the output for considerably more than they pay for the incremental inputs. Most economists have been content to invoke the idea of cyclical fluctuations in productivity in thinking about this fact. My point in this paper is that the fact almost certainly involves a dramatic failure of the principle that marginal cost is equated to price. Marginal cost is literally the increase in the cost of inputs needed to produce added output. That increase is small, so marginal cost is small. When it is compared to price, a large gap is found in most industries. The most obvious explanation of the finding of price far in excess of marginal cost is

monopoly power in the product market. For a straightforward profit maximizing monopoly, the parameter β^* could be interpreted as the reciprocal of the elasticity of demand. That elasticity is around four for the typical manufacturing firm, according to the results. Since few American firms are simple monopolies, the finding probably requires a more elaborate interpretation in terms of theories of oligopoly and product differentiation. Then the finding lends strong support to the view that these theories are more realistic than the simple theory of competition.

Departures from competition in the product market are not the only potential explanation of the finding of this paper. Monopsony in input markets is another possibility. For example, a monopsonist in the labor market faces a marginal cost of labor in excess of the wage it pays. In principle, a firm with sufficient monopsony power in the labor market but facing competitive conditions in its product market could have its price equal to its actual marginal cost, but well above the level inferred from the quoted wage in my calculations. However, I am not aware of any reason to think that monopsony in input markets is anywhere near pervasive enough to explain these findings. On the other hand, simple monopoly or more complicated types of monopoly power in labor or other input markets have no role in explaining the finding. In the labor market, all that is needed for my purposes is that the measured wage is the actual incremental cost of labor. Broader efficiency issues will rest on the question of whether the wage correctly values the foregone time of workers, but the narrow hypothesis that the firm is a price-taker in input markets is all that is needed for measuring the price-

marginal cost ratio.

A significant issue of interpretation arises in the case of a firm that purchases inputs under contracts. Contracts have received the most attention in the labor market, though they could distort the measurement of marginal cost in the case of any input. A contract that predetermines the actual incremental price paid by the purchaser of an input will not distort the calculations I have made, unless the contract price is not correctly measured in the wage or input price data. Contracts of this type create no more than a data problem, and one that is probably not too severe. A conventional commercial contract that specifies both price and quantity is not a likely source of distortion either. Under such a contract, the allocational price is the current market price. Only if the price data involve averages over contracts negotiated in the past will the calculations go astray. Under such contracts, all that is needed is to take total quantities of each input, under contract or otherwise, and compute the revenue share by using the current market price. Even when these calculations fail to use the market price, the bias in the estimate of β is likely to be small and negative, for the reasons discussed in the previous section.

The previous section considered contracts that set input levels by a mechanism other than equating the marginal benefit of the input to the quoted price. The analysis shows that contracts of that type will not distort the estimate of β as long as the average share of the input is correctly measured. This gives considerable assurance that contracts for inputs are not the explanation for the finding of price far above marginal cost.

Errors in measuring labor input are probably the most likely source of bias in the estimates of β. Purely random errors in measuring hours of work do not generate a bias; the errors must be correlated with the right-hand variable, the change in capacity utilization. My data take advantage of all available data to offset this problem; they rely on employer data in the case of workers paid by the hour, and hours reported by salaried workers in the Current Population Survey. Further, the estimates of β seem to be the highest in those industries, such as paper, where most workers are classified as production workers and are paid by the hour.

Fluctuations in effort per hour of work, occurring under a wage-smoothing contract, are another potential explanation of part of the finding that measured productivity growth is strongly correlated with output growth. In the absence of direct measures of work effort, it is difficult to measure the importance of fluctuations in effort. In my judgment, it is unlikely that such fluctuations can fully explain the findings of this paper.

Changes in the utilization of capital services have been offered as an explanation of procyclical productivity growth. If the fluctuations occur through additional hours of homogeneous labor, Solow's productivity calculation takes them properly into account, but if the incremental labor is paid more, then a slight error occurs. The error makes productivity growth positively correlated with output growth, but the correlation is probably much too small to account for the finidings reported here.

Increasing returns to scale could explain some part of the positive estimate of β, but a specification that permits increasing

returns did not to any important extent reverse the findings of strongly positive values for β.

References

Martin Neil Baily, "Wages and Employment under Uncertain Demand," *Review of Economic Studies* **41**:37-50, January 1974

Mark Bils, "Cyclical Behavior of Marginal Cost and Price," manuscript, July 1985

Robert E. Lucas, "Capacity, Overtime, and Empirical Production Functions, " *American Economic Review Papers and Proceedings* **60**:23-27, May 1970

James N.Rosse, "Estimating Cost Function Parameters without Using Cost Data," *Econometrica* **38**:256-275, March 1970

Thomas J. Sargent and Neil Wallace, "The Elasticity of Substitution and Cyclical Behavior of Productivity, Wages, and Labor's Share," *American Economic Review Papers and Proceedings* **64**:257-263, May 1974

Robert M. Solow, "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics* **39**:312-320, August 1957