INCREASING RETURNS: THEORY AND MEASUREMENT WITH INDUSTRY DATA

Robert E. Hall

Hoover Institution and Department of Economics, Stanford University

National Bureau of Economic Research

September 13, 1988

This research is part of the Program in Economic Fluctuations of the National Bureau of Economic Research and was supported in part by the National Science Foundation. I thank David Bizer, Michael Knetter, and James Conklin for superlative research assistance.

Introduction

Earlier research of mine [1988] has reached the conclusion that the firms in some industries have market power. The evidence is simple. When output rises, cost rises by less than price times the change in quantity. Marginal cost is less than price, which is my definition of market power. The procyclical behavior of the Solow productivity residual is the basic fact underlying the research.

One of the explanations of the earlier findings is that firms have monopoly positions, achieved through government restrictions on entry or by other means. In that case, the earnings of the firms should include an element of monopoly profit as well as the normal return to capital. A second explanation is that entry is free, but firms incur fixed costs in order to operate. Then the equilibrium will involve just enough market power to cover fixed costs. Earnings will not exceed the normal return to capital.

A very simple strategy can tell the two explanations apart. Under constant returns to scale (no fixed costs), the telltale cyclical behavior of the Solow residual should disappear once a simple correction is made in the computation of the residual. The correction is to measure labor's share in relation to cost rather than revenue. Because cost will be lower than revenue, the cost-based share will exceed the revenue-based share; the cyclicality of the Solow residual will vanish once a higher share is applied to labor growth. On the other hand, with fixed costs and free entry, revenue and cost will be the same, so the cost-based Solow residual will have the same cyclical behavior as the original revenue-based one. In other words, the research involves the repetition of the tests of my earlier paper, replacing the revenue-based Solow residual with the cost-based residual. The data show that industries in the U.S. rarely have much profit beyond the normal return to capital. Hence the cost-based Solow residual looks pretty much like the revenue-based one, and the empirical results in this paper are much like the ones in my earlier paper.

The paper also derives a method for estimating an index of returns to scale; the index is the elasticity of output with respect to total input. Estimates of the index exceed three in quite a number of industries. Fixed costs or other types of increasing returns appear to be an important feature of some industries.

1. Productivity measurement with market power

Suppose output, Q, is produced by capital, K, and labor, N, in accord with the production function, F:

$$Q = F(K,N) . \tag{1.1}$$

A good approximation to the change in output (as in Solow [1957]) is

$$\frac{\Delta Q}{Q} \doteq \frac{K}{Q}\frac{\partial F}{\partial K}\frac{\Delta K}{K} + \frac{N}{Q}\frac{\partial F}{\partial N}\frac{\Delta N}{N} . \tag{1.2}$$

Define lower-case letters as the logarithms of the corresponding upper-case variables. Then

$$\Delta q \doteq \frac{K}{Q}\frac{\partial F}{\partial K}\Delta k + \frac{N}{Q}\frac{\partial F}{\partial N}\Delta n . \tag{1.3}$$

Let the firm be a price-taker in the capital services market at rental price r and in the labor market at wage w. Conditions for the minimization of cost are

$$\frac{\partial F}{\partial K} = \frac{r}{\lambda} \text{ and } \frac{\partial F}{\partial N} = \frac{w}{\lambda} . \tag{1.4}$$

Here λ is a Lagrangian interpreted as marginal cost. Now define the returns-to-scale function, $\gamma(K,N)$, from

$$K\frac{\partial F}{\partial K} + N\frac{\partial F}{\partial N} = \gamma(K,N)F(K,N) . \tag{1.5}$$

With constant returns to scale, $\gamma(K,N) = 1$; with increasing returns, $\gamma(K,N)$ will exceed one. Now solve for the marginal products:

$$\frac{\partial F}{\partial N} = \frac{w\gamma Q}{rk + wN} = \alpha\gamma\frac{Q}{N} ; \tag{1.6}$$

$$\frac{\partial F}{\partial K} = \frac{r\gamma Q}{rk + wN} = (1 - \alpha)\gamma\frac{Q}{N} . \tag{1.7}$$

Here α is the share of labor cost, wN, in total factor cost, $wN + rK$. Note the important difference between the approach taken here and the one in my earlier paper, Hall [1988]. The earlier work followed Solow in measuring the elasticity of output with respect to labor input by labor's share in revenue. That measure is appropriate under competition. In this paper, I measure the elasticity of output with respect to labor as labor's share in cost multiplied by the returns-to-scale index. No assumption of competition is required.

I can now state the basic relationship studied in the paper:

$$\Delta q = \gamma[\alpha\Delta n + (1-\alpha)\Delta k] \quad . \tag{1.8}$$

The percent change in output is the weighted percent changes in inputs, multiplied by the returns-to-scale index, γ. The weights for the inputs are the corresponding cost shares, α and $1-\alpha$. If γ is roughly a constant, then equation 1.8 is an estimating equation: γ can be estimated as the ratio of the actual change in output, Δq, to the amount by which output should change under constant returns, $\alpha\Delta n+(1-\alpha)\Delta k$, when some exogenous event changes product demand or factor supplies.

Increasing returns also provides some part of the explanation for the well-known positive correlation of productivity and output. Consider the variant of Solow's total factor productivity measure obtained by using cost shares in place of revenue shares. As a practical matter, it makes almost no difference whether cost or revenue shares appear in the productivity measure, because pure profit is sufficiently small that cost and revenue are almost the same. The cost-based Solow productivity residual is

$$\Delta q-\alpha\Delta n-(1-\alpha)\Delta k = (\gamma-1)[\alpha\Delta n + (1-\alpha)\Delta k] \quad . \tag{1.9}$$

Under constant returns to scale ($\gamma=1$), the cost-based Solow residual has the crucial property of *invariance*: When an exogenous event perturbs output and inputs, the residual records no change in productivity. Only true shifts of the production function make the residual depart from zero. On the other hand, with increasing returns ($\gamma > 1$), the residual is positive when output rises, even when there has been no shift of the production function. The residual confuses increases in scale with shifts of the production function. If the value of γ were known, an adjusted Solow residual could be calculated in order to measure the shift in the production function. The strategy used in this paper is the opposite. By examining changes in output associated with events known not to shift the production function, it is possible to estimate γ.

The cost-based Solow residual has the important property that it measures the shift of the production function correctly in the presence of market power. Solow's original approach has the disadvantage of recording false movements of the production function for firms with market power, even with constant returns to scale. When revenue exceeds cost, because of pure monopoly profit, the revenue share of labor understates the elasticity of output with respect to labor input. When some exogenous event raises labor input relative to capital input, the revenue-based Solow residual fails to account for all of the increase in output, because it gives too little weight to labor. My earlier work exploited this property to diagnose and measure market power.

The estimating equation of my earlier work is

$$\Delta q = \mu\alpha\Delta n + (1 - \mu\alpha)\Delta k \ . \tag{1.10}$$

The parameter μ measures the extent of market power; when μ exceeds one, it raises the weight given to labor input above its revenue share, α, and lowers the weight given to capital input. Constant returns to scale is a maintained assumption, so the weights always sum to one. By contrast, the estimating equation in this paper is

$$\Delta q = \gamma[\alpha\Delta n + (1 - \alpha)\Delta k] \quad . \tag{1.11}$$

The effect of the parameter γ is to raise the weights given to both factors, which is appropriate when there are increasing returns to scale. Note that failure of increasing returns is a substantial specification error in the estimating equation for μ. If γ exceeds one, then it is highly likely that the estimate of μ will also exceed one. However, the conclusion of my earlier research—that pure competition in the sense of price equal to marginal cost does not prevail in quite a number of industries—is not damaged by the findings of this paper. Price cannot equal marginal cost with increasing returns to scale.

Technical change and stochastic investment

The previous derivations assumed explicitly that there were no true shifts of the production function over time and implicitly that there was an observed rental cost of capital, r, to which firms equated the marginal product of capital at all times. Neither of these assumptions is realistic. As a step toward realism, suppose that the production function shifts over time in accord with a Hicks-neutral index, Θ:

$$Q = \Theta F(K,N) . \tag{1.12}$$

Let θ be the proportional growth in Θ: $\theta = \Delta\Theta/\Theta$. I will think of θ as a random variable with a positive mean. Let α be the true cost share of labor, where capital cost is measured as the shadow cost of capital (capital's realized marginal product). Then

$$\Delta q = \gamma[\alpha\Delta n + (1 - \alpha)\Delta k] + \theta . \tag{1.13}$$

However, α is unobserved. It has an observed counterpart,

$$\tilde{\alpha} = \frac{wN}{rk + wN} . \tag{1.14}$$

Here r is an observed cost of capital containing a random expectation error, ϵ. The expectation error arises from lags in the investment process. The quantity of capital is set in advance, based on expectations of the demand schedule facing the firm and the interest rate and other determinants of the rental price of capital. The realized marginal product of capital differs by ϵ. Although ϵ has rational expectations properties, these cannot be exploited in this research, because ϵ and θ appear together.

Some algebra shows that the difference between the true labor share and the observed one is

$$\alpha - \tilde{\alpha} = (1 - \tilde{\alpha})\alpha\epsilon . \tag{1.15}$$

Then the estimating equation is

$$\Delta q = \gamma[\tilde{\alpha}\Delta n + (1 - \tilde{\alpha})\Delta k] + \gamma(1 - \tilde{\alpha})\alpha\epsilon(\Delta n - \Delta k) + \theta \qquad (1.16)$$

The factor $\gamma(1-\tilde{\alpha})\alpha$ is close to a constant. Suppose there is an instrumental variable that is a candidate for identifying the estimating equation—its movements do not cause changes in true productivity, θ, and changes in productivity do not cause its movements. Such an instrument is certainly correlated with the expectation error, ϵ, and with the change in the labor-capital ratio, $\Delta n - \Delta k$. Nonetheless, it is a reasonable identifying assumption that an instrument is uncorrelated with the product, $\epsilon(\Delta n - \Delta k)$. The product is positive in good times (when both its factors are positive) and is positive as well in bad times (when both its factors are negative). The instrument will be positive in good times and negative in bad times. Hence its correlation with the product will be close to zero. More generally, if the three random variables ϵ, $\Delta n - \Delta k$, and the instrument have a symmetric joint distribution, the correlation will be exactly zero, because the correlation is a third moment.

The test of the null hypothesis of constant returns, $\gamma = 1$, can be done in the standard estimation framework. Note that the null hypothesis assigns a fixed value to γ even though the assumption that γ is fixed is restrictive in a broader class of technologies. An alternative approach to hypothesis-testing brings out the close connection of this research to Solow's productivity measure. By subtracting the weighted growth of factor input from both sides of equation 1.16, the left-hand side becomes Solow's residual:

$$\Delta q - \tilde{\alpha}\Delta n - (1 - \tilde{\alpha})\Delta k =$$
$$(\gamma - 1)[\tilde{\alpha}\Delta n + (1 - \tilde{\alpha})\Delta k] + \gamma(1 - \tilde{\alpha})\alpha\epsilon(\Delta n - \Delta k) + \theta \quad (1.17)$$

With constant returns to scale, the first term on the right-hand side disappears. I have already argued that the other two terms should be uncorrelated with a properly chosen instrument. Consequently, the following result is established:

Invariance Theorem. *Under constant returns to scale, the cost-based Solow residual is uncorrelated with an instrumental variable, irrespective of the amount of market power.*

The empirical results presented in this paper are of two types. The first shows the failure of the invariance property by constructing cost-based Solow residuals and showing that they are positively correlated with instruments that themselves are positively correlated with output. The second measures the extent of increasing returns by estimating γ as a parameter.

2. Econometric method and choice of instruments

The invariance proposition tested in this paper is similar in form to the one tested in my earlier paper, Hall [1988]. The null hypothesis is refuted by finding a positive correlation between the productivity residual and an exogenous instrument. Econometrically, the simplest way to test for the absence of correlation is to calculate the regression coefficient of the productivity residual on the instrument and use the t-test for inference.

To be useful as an instrument, a variable must be the cause of important movements in the output and employment of an industry, but not a cause or an effect of shifts in its productivity. Here I use the same three instruments as in my previous work: the rate of growth of military spending, the rate of change of the price of crude oil, and the political party of the President. All are correlated with the output and employment of at least some of the industries studied here. For a more extensive defense of their exogeneity with respect to random productivity shifts, see my earlier paper.

3. Data

Most of the data used in this study are the same as described in my earlier paper (Hall [1988]). These include real value added, compensation, and total hours of work, and the real capital stock. The only series used here that was not part of the earlier work is the rental price of capital.

Construction of the rental price follows Hall and Jorgenson [1967]. The formula relating the rental price to its determinants is:

$$r = (\rho + \delta) \frac{1 - k - \tau d}{1 - \tau} p_K \ . \qquad (3.1)$$

The determinants are:

ρ: The firm's real cost of funds, measured as the dividend yield of the S&P 500 portfolio;

δ: The economic rate of depreciation, 0.127, obtained from Jorgenson and Sullivan [1981], Table 1, p. 179;

k: The effective rate of the investment tax credit, from Jorgenson and Sullivan, Table 10, p. 194;

d: The present discounted value of tax deductions for depreciation, from Jorgenson and Sullivan, Table 6, pp. 188-189;

p_K: The deflator for business fixed investment from the U.S. National Income and Product Accounts.

Use of the dividend yield as the real cost of funds is justified by two considerations: First, the great bulk of investment is financed through equity in the form of retained earnings. Second, the use of a market-determined real rate avoids the very substantial problems of deriving an estimated real rate by subtracting expected inflation from a nominal rate. The dividend yield is a good estimate of the real cost of equity funds whenever the path of future dividends is expected to be proportional to the price of capital goods. For the typical firm, this is an eminently reasonable hypothesis. Of course, for firms with low current dividend payouts and high expected growth, the dividend yield understates the real cost of funds. But these firms are counterbalanced by mature firms whose payouts are high and whose growth rates are below the rate of inflation.

4. Results

Table 1 shows the basic data for nondurables manufacturing. The first column is the rate of growth of output; the second is the rate of growth of hours of work, the thrid is capital stock growth, and the fourth is labor's share in total cost. The cost-based residual in the fifth column is obtained by multiplying hours growth by the labor share, multiplying capital growth by one minus the labor share, and subtracting the sum from output growth. The last two columns show the values of two instruments—the rate of growth of the price of crude oil and the rate of growth of military spending. There is a noticeable negative correlation between each of the instruments and the growth of output, one the one hand, and the cost-based residual, on the other hand. Oil price increases in 1957, 1973-75, and 1978 were associated with low or negative rates of growth of output and measured productivity. Oil price declines in 1959, 1963-65, and 1972 were coupled with high output growth and large measured productivity residuals. For military spending, increases in 1966-67 came at the same time as low or negative growth rates of output and measured productivity. Declines in military spending in 1955 and 1971-73 coincided with high measured productivity growth. The evidence based on military spending is more mixed; for example, in 1954, a large decline in military spending was associated with a decline in output but measured productivity growth was only slightly below normal.

Table 1: Cost-based residual (percent or percent change) for non-durables

Year	Output Growth	Hours Growth	Capital Growth	Labor Share	Cost-based Residual	Instruments: Oil	Instruments: Military
1953	3.6	1.8	2.3	78.4	2.2	7.0	4.8
1954	-0.7	-4.3	2.0	79.6	2.7	2.9	-8.3
1955	7.3	3.7	1.2	80.9	4.4	0.1	-17.0
1956	3.4	0.6	3.0	80.1	2.9	0.5	-6.1
1957	0.4	-1.9	3.0	78.7	1.9	9.7	2.5
1958	-0.3	-4.5	0.5	79.1	3.3	0.2	-2.6
1959	9.0	4.5	-0.1	80.5	5.3	-3.4	-1.2
1960	1.0	-0.4	1.5	80.4	1.2	-0.6	-2.0
1961	2.3	-0.7	1.9	81.1	2.9	0.3	1.6
1962	5.8	2.4	2.3	82.0	3.8	0.2	3.5
1963	7.0	0.1	2.5	82.2	7.0	-0.4	-2.6
1964	4.7	1.2	3.4	82.8	3.7	-0.4	0.9
1965	5.5	3.3	6.3	82.6	2.8	-0.1	-9.2
1966	5.5	3.9	7.7	81.4	2.4	0.7	8.7
1967	-0.5	0.1	5.8	81.1	-0.5	1.1	17.2
1968	5.8	1.8	4.7	80.6	4.4	0.8	8.0
1969	3.4	1.2	4.6	78.9	2.4	4.3	-4.6
1970	-0.3	-3.7	4.1	77.2	2.6	0.9	-7.6
1971	3.9	-2.2	2.8	78.6	5.6	7.7	-10.7
1972	6.8	2.0	2.6	79.4	5.2	-0.7	-6.0
1973	8.3	2.0	2.6	79.3	6.7	10.2	-8.6
1974	-6.8	-2.9	4.5	76.6	-4.6	51.9	-7.0
1975	-2.6	-6.9	3.4	75.0	2.6	14.8	-2.2
1976	9.0	4.7	3.3	76.5	5.4	3.2	-4.0
1977	6.3	2.1	3.3	76.0	4.7	7.8	0.5
1978	3.3	1.9	3.3	74.6	1.9	9.0	0.5
1979	2.2	0.3	3.2	74.0	2.0	22.7	2.6
1980	-3.6	-2.8	2.7	73.0	-1.6	39.1	3.0

The regressions to carry out formal tests of the invariance of the cost-based productivity residual in nondurables are:

$$\Delta q - \tilde{\alpha}\Delta n - (1-\tilde{\alpha})\Delta k = \underset{(.004)}{.0327} - \underset{(.030)}{0.143}\, x_{OIL} \quad \text{DW: } 1.78 \qquad (4.2)$$

$$\Delta q - \tilde{\alpha}\Delta n - (1-\tilde{\alpha})\Delta k = \underset{(.005)}{.0212} - \underset{(.070)}{0.114}\, x_{MIL} \quad \text{DW: } 1.99 \qquad (4.3)$$

In both cases, the correlation of the instrument with output growth is negative, so the evidence is unambiguous that an event such as an oil price decline or cut in military spending, which stimulates nondurables sales, raises the cost-based measure of productivity. The failure of the theoretical invariance property is attributable to some failure of its underlying assumptions. My primary interpretation is that constant returns to scale fails. I will return later to a fuller discussion of the implications of the rejection of the invariance proposition.

Tables 2 and 3 present the evidence for one- and two-digit industries and for the three instruments. The entries in the tables are the marginal significance levels for the invariance test. That is, each number is the probability that a covariance at least as positive as the one found might have arisen purely by chance. In Table 2, for one-digit industries, there is at least one instrument that gives rejection at the 10-percent level for each industry. The military spending instrument gives fairly strong evidence of increasing returns in nondurables and services. The oil price

Table 2: Marginal significance levels for one-digit industries

Industry	Military Spending	Oil Price	Political Party
Construction	0.546	0.070	0.090
Durable goods	0.272	0.018	0.175
Nondurable goods	0.058	0.000	0.291
Transportation & public utilities	0.119	0.007	0.274
Trade	0.265	0.001	0.580
Finance, insurance, & real estate	0.218	0.842	0.028
Services	0.040	0.769	0.092

Table 3: Marginal significance levels--further industry detail

Industry	Military Spending	Oil Price	Political Party
20: Food & kindred products	0.341	0.013	0.300
21: Tobacco manufactures	0.334	0.268	0.242
22: Textile mill products	0.254	0.105	0.158
23: Apparel & other textile products	0.269	0.620	0.591
24: Lumber & wood products	0.480	0.527	0.129
25: Furniture & fixtures	0.036	0.126	0.486
26: Paper & allied products	0.073	0.008	0.367
27: Printing & publishing	0.136	0.013	0.440
28: Chemicals & allied products	0.104	0.000	0.392
29: Petroleum & coal products	0.114	0.001	0.533
30: Rubber & misc. plastic products	0.607	0.052	0.135
31: Leather & leather products	0.160	0.631	0.140
32: Stone, clay, & glass products	0.177	0.005	0.565
33: Primary metal industries	0.272	0.369	0.233
34: Fabricated metal products	0.075	0.001	0.143
35: Machinery, except electrical	0.311	0.041	0.242
36: Electrical & electronic equip.	0.309	0.022	0.033
38: Instruments & related products	0.494	0.124	0.312
39: Miscellaneous manufacturing	0.195	0.001	0.076
48: Communication	0.341	0.144	0.665
49: Elec., gas, & sanitary services	0.457	0.060	0.586
371: Motor vehicles	0.285	0.112	0.383
372-9: Other transportation equip.	0.688	0.024	0.441

Notes:

Marginal significance levels for a one-tailed test of the hypothesis that the covariance of the cost residual and the instrument is positive. The sign of the instrument is normalized so that its covariance with output growth is positive.

instrument gives quite strong evidence in five of the seven industries. The oil price is a factor price and a source of demand shifts for each of the industries. On the assumption that neither role should shift the production function, that is, that oil price fluctuations are uncorrelated with true productivity growth, I conclude that invariance fails in the direction predicted by increasing returns to scale.

In three industries—construction, finance-insurance-real estate, and services—the political party of the President gives reasonably strong rejection of invariance. The political party variable is suitable as an instrument to the extent that the differences in policies of the two parties create differences in output growth rates but not in true productivity growth. In fact, real growth has generally been greater under Democrats than under Republicans. Under the assumption that the growth was achieved through differences in monetary and fiscal policy and not through differences in policies affecting the production function, the political dummy is a good instrument. I consider this assumption eminently reasonable.

The character of the more detailed results in Table 3 is similar. Failure of invariance of the productivity measure is most common for the oil price instrument. Indeterminate results arise mainly because output growth has not been correlated with the instrument.

Estimates of the returns-to-scale index

Estimates of γ obtained by applying instrumental variables to equation 1.16 often have very large standard errors. High apparent dispersion will occur whenever an instrument is strongly correlated with output but weakly correlated with weighted factor input. The high dispersion does not convey any uncertainty about the failure of constant returns—that hypothesis would require that the covariances of output and weighted input with the instrument be the same. Rather, the uncertainty is over how much greater than unity is γ. A more informative procedure is to estimate the reciprocal, γ^{-1}. By mapping all values of γ greater than one into the interval between zero and one, the procedure of estimating the reciprocal gives a much more interpretable estimate of the sampling variation of the estimate of γ.

Tables 4 and 5 present the evidence in the form of estimates of γ. Estimation is by two-stage least squares, using all three instruments together. In both tables, the first column provides the estimates of γ^{-1} and their standard errors. The second column gives the Durbin-Watson statistics for the estimates. The third column provides the corresponding estimate of γ. Although the reciprocal of the two-stage least squares estimator of a coefficient is not exactly the two-stage least squares estimator of the reciprocal of the coefficient, in practice the values in the third columns of the two tables are close to the results of the application of two-stage least squares directly to equation 1.16.

Table 4: Estimates of returns-to-scale index at the one-digit level

Industry	Estimate of reciprocal of index	Durbin-Watson statistic	Reciprocal of estimate
Construction	0.597 (0.277)	1.091	1.675
Durable goods	0.543 (0.128)	1.826	1.841
Nondurable goods	0.322 (0.110)	1.990	3.107
Transportation & public utilities	0.100 (0.169)	1.337	10.030
Trade	0.224 (0.178)	2.390	4.468
Finance, insurance, real estate	0.353 (0.250)	1.001	2.830
Services	0.926 (0.220)	2.505	1.080

Note: Standard errors in parenthesis.

Instruments:
Defense expenditures, price of oil, and political party.

Table 5: Estimates of returns-to-scale index: further industry detail

Industry	Estimate of reciprocal of index	Durbin-Watson statistic	Reciprocal of estimate
20: Food and kindred products	0.030 (0.132)	1.437	33.557
21: Tobacco manufactures	0.256 (0.358)	2.032	3.909
22: Textile mill products	0.500 (0.152)	1.920	1.999
23: Apparel and other textile products	0.933 (0.271)	2.080	1.072
24: Lumber and wood products	0.725 (0.276)	2.005	1.379
25: Furniture and fixtures	0.736 (0.141)	2.340	1.359
26: Paper and allied products	0.208 (0.079)	1.582	4.810
27: Printing and publishing	0.384 (0.165)	1.264	2.605
28: Chemicals and allied products	0.007 (0.091)	1.218	138.889
29: Petroleum and coal products	-0.309 (0.187)	1.292	-3.236
30: Rubber and miscellaneous plastic products	0.606 (0.144)	2.323	1.650
31: Leather and leather products	0.212 (0.238)	1.782	4.710
32: Stone, clay, and glass products	0.461 (0.109)	1.807	2.170

33: Primary metal industries	0.351 (0.117)	1.909	2.852
34: Fabricated metal products	0.230 (0.249)	2.247	4.352
35: Machinery, except electrical	0.681 (0.168)	1.920	1.469
36: Electric and electronic equipment	0.447 (0.166)	2.483	2.237
38: Instruments and related products	0.474 (0.358)	2.098	2.111
39: Miscellaneous manufacturing industries	0.182 (0.156)	2.237	5.491
48: Communication	0.834 (0.736)	2.223	1.199
49: Electric, gas, and sanitary services	0.496 (0.206)	2.141	2.016
371: Motor vehicles and equipment	0.382 (0.189)	3.370	2.621
372-9: Other transportation equipment	0.886 (0.123)	1.790	1.129

Note: Standard errors in parenthesis.

Instruments: defense expenditures, price of oil, and political party.

Tables 4 and 5 show that the failure of constant returns in many industries is quite profound. The estimated elasticity of output with respect to total input, γ, is above 1.5 in all one-digit industries in Table 4 save services. In three industries—nondurables, transportation-utilities, and trade—output rises by more than three percent when an outside force makes input rise by one percent. In Table 5, 7 of the 23 industries have return-to-scale indexes of greater than three.

5. Interpretation of rejection of the invariance hypothesis

As a matter of theory, an optimizing firm with power in its output market and a constant-returns technology should obey invariance—its cost-based productivity residual should be uncorrelated with any outside force that changes output but does not shift its production function. I have shown earlier that increasing returns could explain the failure of invariance in the direction found here. The productivity residual uses the cost share to measure the elasticity of output with respect to labor input. In the presence of fixed costs or other failures of constant returns, the cost share understates the true elasticity. Then, as a result of the understatement, the cost-based residual would incorporate too small an adjustment for variations in labor input and the residual itself would rise every time output rose.

Although my discussion has considered increasing returns in the firm's own technology, the invariance proposition will also fail when the firm's technology has constant returns but there is an externality that makes one firm's output complementary with other firms' output. The thick-market externality discussed by Diamond [1982] is a leading example. The extremely uneven geographical distribution of economic activity suggests that thick-market effects are strong. Efficiency is greater in places and at times when suppliers, workers, and customers are dense. The overall technology of an industry with a thick-market externality will have increasing returns even though each firm has constant returns.

Thick-market effects internal to the firm may be an important source of increasing returns for the firm, as well. Consider a package delivery service. When its customers become more numerous, its operations become more efficient because each truck can make more stops in each area, and deliver more packages per mile of driving.

Other conditions besides increasing returns could cause the failure of invariance of the productivity residual. These include chronic excess capacity, unmeasured fluctuations in work effort, suboptimal levels of employment because of monopsony power in the labor market, and unmeasured fluctuations in capital utilization.

Labor hoarding

Before considering the various specification errors in turn, I should discuss one major phenomenon that has an important role in the story told by the data but is *not* an alternative explanation of the findings. I refer to labor hoarding and overhead labor. The following example shows how the invariance property of the cost-based residual holds in the presence of overhead labor:

Suppose that the technology is such that the level of employment required to produce output Q is $\lambda K+\phi\, max(Q-K,0)$. That is, with a capital stock of K and overhead labor of λK, it is possible to produce up to K units of output. Additional output requires an increment of ϕ units of labor for each unit of output above K. The shadow value of capital is $-\lambda w$ when output is below K because the firm could produce just as much output with lower overhead labor if its capital were lower. The shadow value of capital is $(\phi-\lambda)w$ when Q exceeds K—in that regime, more capital requires more overhead workers but reduces the requirement for the incremental labor described by ϕ. Let β be the probability that output will exceed K. Then the expected shadow value of capital is $(\beta\phi-\lambda)w$. At the optimum capital stock, the expected shadow value of capital equals the service price of capital, r. Hence, $\beta=(r+\lambda w)/\phi w$. Suppose that the fluctuations in output are in a small region above and below K. The cost share, α, will be close to $w\lambda/(w\lambda+r)$. Because there is no true productivity change, the actual change in output, Δq, is a valid instrument itself. Suppose that the capital stock does not change over time. When output is below K, the change in employment is zero and the cost-based residual is equal to Δq. Thus the relation between the residual and the instrument has a unit slope. When output is above K, the change in employment, ΔN, is $\phi\Delta Q$. The level of employment is close to λQ, because Q is only a little over K. Hence, the rate of growth of employment, Δn, is approximately $(\phi/\lambda)\Delta q$. The slope of the relation between the cost-based residual $\Delta q-\alpha\Delta n$ and Δq is $1-\alpha\phi/\lambda$. The average slope is $1-\beta+\beta(1-\alpha\phi/\lambda)$. Inserting the values for β and α derived above shows that the average slope is zero.

In the example, it is true that when the firm is in the labor-hoarding regime (Q is below K), the covariance of the cost-based residual and the instrument would be strongly positive. However, this is exactly counterbalanced by a negative covariance when output is above K. What if a firm spent most of its time in the labor-hoarding regime and had output above K only in times of extreme demand? Isn't this the normal case for most firms? The answer is that such a firm is not satisfying the condition for optimal investment; it has excess capacity.

Labor hoarding and overhead labor are probably important phenomena in a number, if not the majority, of the industries studied in this paper. When a firm is in a labor-hoarding regime, its cost-based residual will be positively correlated with an instrument. In that respect, labor hoarding is an essential part of the explanation of the findings of this paper. However, labor hoarding is not an alternative explanation of the failure of the invariance property. Fixed costs or other types of increasing returns are likely to underlie chronic operation in a labor-hoarding regime. A firm with a constant returns technology and an optimal investment strategy, no matter how ridden with forecasting errors, will spend enough time in a labor-shortage regime to offset the time spent in the labor-hoarding regime. As the example shows, the condition for optimal investment amounts to stating that the two regimes combine in such a way as to eliminate any covariance of the cost-based residual with an instrument.

Chronic excess capacity

If firms consistently hold more than the optimal amount of capital, invariance of the cost-based residual will fail. With chronic excess capacity, the firm's costs would be higher than appropriate, so the cost share of labor would understate the true elasticity of output with respect to labor input. Some theories of the strategic interaction of firms have suggested the desirability of capacity above the cost-minimizing level for the realized distribution of output. Excess capacity makes credible a threat to revert to competition. However, it seems unlikely that the conditions for this motive for holding excess capacity are widespread and powerful enough to explain the findings of this paper.

Unmeasured fluctuations in work effort

Of the various specification errors that may have biased the covariance of the cost-based residual and an instrument upward, the most important is the following, considered at length in my earlier work: There are unmeasured variations in work effort that are positively correlated with output. When an outside force drives up output and employment, measured productivity rises for a reason unrelated to increasing returns. There is no question that the method of this paper is vulnerable to such measurement errors; the only question is the numerical importance of the errors.

A number of considerations convince me that unmeasured fluctuations in effort cannot explain all of the correlation I find between the cost-based residual and various instruments. First, the magnitude of the fluctuations would have to be large. My earlier work showed that the effort of the typical workers would have to have been almost 10 percent above normal for a sustained period in the 1960s, for example. Second, survey evidence collected from employers by Fay and Medoff [1985] suggests that effort is slightly negatively correlated with output, not strongly positively, as required to give an upward bias in the estimated returns to scale index. Third, the fluctuations in effort needed to rationalize the observed fluctuations in productivity are inconsistent with the observed behavior of compensation. Work effort rises so much in a boom that the wage, corrected for changes in effort, actually falls. I find this implausible. The only way to rescue the hypothesis of large fluctuations in work effort is to invoke the theory of wage smoothing, in which workers are not paid on a current basis for their labor input, but rather receive compensation based on the average level of work over an extended period.

Other labor issues

A basic maintained hypothesis of this paper is that the firm chooses an optimal level of employment. The derivation of equation 1.16 makes the assumption that the marginal revenue product of labor is equated to the wage. An alternative is that the firm employs too few workers, on the average. Then the measured cost share of labor would understate the true elasticity because of the understatement of effective labor cost, and the covariance of the cost-based residual and an instrument would be explained. For example, if the typical firm has strong monopsony power in its labor market, a failure of the invariance property would occur in the observed direction. But the conditions under which this could be expected to persist for long periods are strenuous. First, if there is bilateral bargaining with a labor union, one would not expect to find a shadow value of labor in excess of the observed wage. Both parties could be made better off by attracting a worker from the open market and paying the worker the prevailing union wage. And if the union has much monopoly power, it is likely to succeed in pushing the observed wage above the shadow value, by extracting a lump-sum component of compensation as part of an efficient bargain.

Second, the firm has a strong incentive to overcome its monopsony position in the labor market by attracting workers from more distant markets. When it can only get more work from its own local market by driving up every worker's wage, it will turn to other markets. What matters is the elasticity of labor supply from the entire labor market to the one firm in the long run. It is hard to believe that this elasticity is anything less than a very large number for most firms.

Mismeasurement of capital

An important implicit assumption of my work is that capital input is correctly measured. The measure of capital I use is the amount of capital available for use. As long as capital has no pure user cost, it is reasonable to assume that all capital available is in use. If there is a pure user cost—if capital depreciates in use rather than just over time—then the situation is different. There is a capital supply decision similar to the labor supply decision and presumably fluctuations in capital input occur in parallel to fluctuations in output. I should note at the outset that if capital is out of use because it is redundant—its shadow value is zero—then there is no bias in my procedure. The dangerous case is when capital has a positive shadow value and there are unmeasured fluctuations in utilization.

Though it is not possible to dispose of this hypothesis as a complete or partial explanation of the failure of invariance, it is possible to show that it calls for rather extreme movements of the true capital stock, corresponding to substantial pure user costs of capital. Let Δv be the change in measurement error of capital actually in use and let $\Delta\tilde{k}$ be the change in measured capital ($\Delta\tilde{k}=\Delta k+\Delta v$). Then the Solow residual, calculated with measured rather than actual capital, under constant returns, will be:

$$\Delta q - \alpha\Delta n - (1-\alpha)\Delta\tilde{k} = -(1-\alpha)\Delta v \quad . \tag{5.1}$$

Because capital measurement errors are likely to be negatively correlated with output changes (an increase in output raises unmeasured capital utilization and lowers v), the errors are likely to contribute to a failure of the invariance condition in the direction found in this paper. For example, suppose that the change in capital measurement error is proportional to the change in labor input per unit of measured capital,

$$\Delta v = -\phi(\Delta n - \Delta\tilde{k}) \quad . \tag{5.2}$$

Strict complementarity of work hours and capital hours would mean that ϕ had the value of one. Then the Solow residual is

$$\Delta q - \alpha\Delta n - (1 - \alpha)\Delta\tilde{k} = \phi(1-\alpha)(\Delta n - \Delta\tilde{k}) \quad . \tag{5.3}$$

Estimation of ϕ by instrumental variables answers the following question: What magnitude of measurement error would be required to explain the observed failure of the invariance condition under constant returns? The answer turns out to be a very large magnitude, well above the intuitive maximum of $\phi=1$. For nondurables, the results of estimation with the three instruments are:

$$\underset{}{\Delta q - \tilde{\alpha}\Delta n - (1-\tilde{\alpha})\Delta k} = \underset{(.012)}{.0549} - \underset{(1.606)}{5.034}\,(1-\alpha)(\Delta n - \Delta\tilde{k}) \quad . \tag{5.4}$$

In order to explain the magnitude of the correlation of the Solow residual with the instruments, the elasticity of the measurement error with respect to the change in labor input must be implausibly large—around 5. The simple model in which capital labor fluctuate in proportion, with $\phi=1$, is not nearly enough to explain the findings of the paper.

6. Conclusions

Under constant returns to scale, with correctly measured data, and cost-minimizing capital stock, the cost-based Solow residual is uncorrelated with any instrument that is uncorrelated with shifts in the technology. This invariance condition fails for a great many U.S. industries. Instead, when an outside force affects product demand or factor supplies, measured productivity changes. In many cases, the magnitude of the shift is dramatic. The elasticity of output with respect to total input is frequently three or more.

Some part of the high elasticity may be the result of measurement errors. Unmeasured shifts in the intensity of work effort, and its counterpart for capital, utilization rates, may explain part of the high measured elasticity of output with respect to total input. However, explanations along this line are not as easy as they might seem. A firm is not minimizing cost if it can command greater work effort or higher utilization at low cost—it should reduce its work force and capital stock once and for all and then set higher effort levels and utilization rates. Similarly, a blanket appeal to labor hoarding cannot survive careful examination as an explanation of the findings.

The construction of aggregate models faithful to these empirical results is a substantial challenge. Sharply increasing returns makes agglomeration of economic activity by location and time highly desirable. It means that a large fraction of the growth in real income arises from agglomeration and not from technical progress. Models with increasing returns, especially those arising from externalities, are likely to have multiple solutions or complete indeterminacy. Increasing returns can overcome the strong tendency of standard neoclassical models to predict smooth evolution of economic activity over time.

References

Peter A. Diamond, "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, vol. 90. pp. 881-894, 1982.

Jon Fay and James Medoff, "Labor and Output over the Business Cycle," *American Economic Review*, vol. 75, pp. 638-655, September 1985.

Robert E. Hall, "The Relation between Price and Marginal Cost in U.S. Industry," *Journal of Political Economy*, vol. 96, October 1988.

Robert E. Hall and Dale W. Jorgenson, "Tax Policy and Investment Behavior," *American Economic Review*, vol. 57, pp. 391-414, June 1967.

Dale W. Jorgenson and Martin A. Sullivan, "Inflation and Corporate Capital Recovery," in *Depreciation, Inflation, and the Taxation of Income from Capital*, Charles R. Hulten (ed.), Washington, DC, Urban Institute, 1981, pp. 171-237.

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